Strategies to deal with ordinal missing data for measurement invariance testing and specification searches – A comparison of commonly used methods

By

Po-Yi Chen

Submitted to the graduate degree program in Department of Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chair: Holger Brandt

_____

Wei Wu

_____

Amber Watts

_____

Kelsie Forbush

_____

Paul Johnson

Date Defended: 8 May 2018

The dissertation committee for Po-Yi Chen certifies that this is the approved version of the following dissertation:

Strategies to deal with ordinal missing data for measurement invariance testing and specification searches – A comparison of commonly used methods

Chair: Holger Brandt

Date Approved: 25 May 2018

Abstract

Measurement equivalent/invariance is a key concept in psychological testing. Failing to correctly identify non-invariant items can lead biased group comparisons and biased selections. The methodological literature on measurement equivalent/invariance (ME/I) and specification searches in structural equation modeling (SEM) usually consider only complete data. In practice, ME/I tests are often done on Likert scales which involve ordinal variables. Missing data on ordinal variables can be problematic in ME/I tests based on the chi-square statistic (χ^2) and modification indices. To deal with missing ordinal data, a recommended strategy is to combine multiple imputation with weighted least squares estimation methods. However, both χ^2 statistic and modification indices are not available with this strategy. Consequently, researchers have to adopt "suboptimal" methods: 1) use full information maximum likelihood (FIML) by treating ordinal data as normally distributed continuous data; 2) use robust FIML by treating ordinal data as non-normally distributed continuous data; and, 3) use weighted least squares (WLSMV) estimators with suboptimal missing data handling techniques, such as pairwise deletion. Previous studies have found that any of the strategies may bias the point estimates and χ^2 statistics in SEM. Yet, there has been no systematic comparison of the suboptimal strategies, especially in the context of ME/I tests or chi-square difference tests ($\Delta\chi^2$ tests). Thus, the goals of my dissertation are to investigate the relative performance of these commonly used suboptimal strategies on the $\Delta\chi^2$ tests and modification indices in ME/I testing with ordinal missing data. Two simulation studies were conducted. Study 1 aimed to compare the three strategies in terms of the accuracy and efficiency of parameter estimates as well as the type I error rate and power of $\Delta\chi^2$ tests. Study 2 aimed to examine the relative performance of the strategies on specification

search. I investigated three backward specification search methods based on the largest modification index using the three suboptimal methods described above and compared it to a recently proposed forward specification search method based on confidence intervals (CI approach), which can be implemented in the "optimal" approach of WLSMV using multiple imputations. The first simulation study showed that when the target data set contains a substantive amount of ordinal missing data, using the $\Delta\chi^2$ tests and modification indices obtained from WLSMV with pairwise deletion lead to a substantive inflation of type I error rates. In contrast, the $\Delta\chi^2$ tests and modification indices obtained from FIML approaches had a better ability to control the type error with sufficient power to test measurement invariance under most conditions. However, parameter estimates were biased for the FIML approaches. In the second simulation study, FIML based modification indices could identify more effectively the correct invariant factor loadings than the modification indices from the WLSMV estimator using pairwise deletion or the CI approach from the WLSMV estimator with multiple imputations. However, all search methods showed an inflated type I error at the model level because none of the methods could effectively locate non-invariant thresholds. Future directions of the ordinal missing data in invariance testing are discussed and practical suggestions for empirical researchers are provided.

Acknowledgments

I would like to thank Drs. Wei Wu, Holger Brandt, Amber Watts, Kelsie Forbush and Paul Johnson for agreeing to be on my dissertation committee. I really appreciate your time and guidance. I thank the Center of Research Methods and Data Analysis and the Department of Psychology at the University of Kansas for providing me resources to conduct some analyses in this dissertation. I would also like to thank Mauricio Garnier-Villarreal, Benjamin Kite, and Fan Jia for discussing with me on the programming and methodological issues related to my dissertation topics. Last, I would like to thank my families for their unconditional support.

Table of Contents

List of Figures

List of Tables

Chapter 1—Introduction

Measurement equivalent/invariance (ME/I) is an important property for psychological tests (Brown, 2006). It concerns whether the relationships among observable indicators and the target latent construct(s) are identical across groups (Millsap, 2012). Multiple group confirmatory factor analysis (MG-CFA) is commonly used to test ME/I. Previous studies have shown that either ordinal or missing data problems can affect the ME/I tests using MG-CFA (e.g., Sass, Schmitt & Marsh, 2014; Widaman, Grimm, Early, Robins, & Conger, 2013). However, very few studies have examined the joint effect of these two problems; that is, the influence of ordinal missing data on ME/I tests with MG-CFA.

ME/I tests typically involve the use of chi-square difference tests and modification indices. To deal with ordinal data, estimators based on polychoric correlations are often recommended (e.g., Flora & Curran, 2004; Sass, et al., 2014), among which the most popular is the so-called weighted least squares with means and variance adjusted (WLSMV) estimator. These estimators do not handle missing data directly. Although multiple imputation may be used with an ordinal data estimator to produce accurate parameter estimates, it is unclear how to pool the chi-squared statistics (χ^2) and modification indices across imputed data (Liu et al., 2017; B. O. Muthén, 2017). This creates a unique and important problem for ME/I tests. For example, if one uses WLSMV with multiple imputations (MI) in Mplus, the average of χ^2s across imputed data sets is usually reported, which will lead to a highly inflated type I error rate for model fit evaluation (Teman, 2012). Similarly, one cannot obtain the modification indices after multiple imputation (B. O. Muthén, 2017).

These limitations greatly hinder ME/I tests and the process of following up model modifications (i.e., specification search). As a result, in order to obtain these crucial statistics, researchers are forced to adopt "suboptimal" strategies for ME/I tests with ordinal missing data. Specifically, they might either (1) treat ordinal data as continuous so that they can use full information likelihood based on normality assumption ($FIML_{mvn}$) or the robust version of it (robust FIML) to handle missing data (e.g., Fokkema, Smits, Kelderman, & Cuijper., 2013), or (2) stay with polychoric correlation based estimators (e.g., WLSMV) to correctly handle the ordinal nature of the data, but use suboptimal missing data methods such as listwise or pairwise deletion (e.g., Zhou, Whealin, Wang & Lee, 2017).

Although previous studies have shown that either strategy could affect the chi-square test statistic and parameter estimates in SEM (e.g., Li, 2016; Marsh, 1998), it is unclear which strategy should be preferred in the context of ME/I testing with ordinal missing data. In this dissertation, I conducted two simulation studies to address the question. Specifically, I examined the relative performances of several suboptimal strategies for ME/I tests and specification searches that researchers might use in practice with the presence of ordinal missing data.

The rest of the dissertation was organized as follows. In Chapter 1, I reviewed background information for the dissertation, including the definition of ME/I (Millsap, 2012), procedures of ME/I testing using MG-CFA, specification searches using MG-CFA, and missing data problems in SEM. I concluded Chapter 1 by summarizing the findings from the reviewed literature and discussing the impacts of ordinal missing data. In Chapter 2, I presented limitations of existing research and raised the research questions for my dissertation. In Chapter 3, I described designs of the two simulation studies.

1.1 Definition of measurement invariance

Conceptually, ME/I indicates that the relationships among observable indicators and target latent constructs are identical across populations/groups. Millsap (2012) provided the math behind the definition of ME/I. Suppose there are k populations, k = 1, 2…… K. Let X = ($X_1,\ldots,X_{p'}$) represent a vector of scores on p' observed indicators; W = ($W_1,\ldots,W_r$) represents a vector of scores on r (all) latent factors underlying X (r < p'). The latent factors in W can be further categorized into two groups W_t and W_n, where W_t are the latent factors measured by X, while W_n are unique factors. Let $P_k(X \mid W_t)$ be the conditional probability of X being in the k_{th} population, which is also called the k_{th} population's measurement response function. One can say that ME/I holds for X in its relation to W_t if and only if

$$P_k(X \mid W_t) = P(X \mid W_t) \quad \forall k = 1 \,.\, K \qquad (1.1)$$

In other words, measurement response functions should be identical across populations under ME/I (Millsap, 2012, p.46).

Note that the definition in Equation 1.1 is based on conditional probabilities, which suggests that different populations could have different levels of W_t (e.g., different means of target constructs) yet have ME/I. However, a problem of this definition is that it is difficult to investigate ME/I empirically, given that W_t are not directly observable. Various methods have

been developed to solve this problem. These methods can be classified as either observed variable or latent variable approaches (Millsap, 2012). For observed variable approaches, one or more measured variables are used as the proxy of W_t (e.g., sum scores of observed indicators). ME/I tests are conducted using methods such as the Mantel-Haenszel method or logistic regression. For the latent variable approach, methods such as item response theory (IRT) models or confirmatory factor analysis (CFA) models are used to directly model W_t. The latent variable approaches are superior to the observed variable approaches because they account for measurement errors. On the other hand, the observed variable approaches usually require a smaller sample size. Several studies have reviewed these methods (e.g., Vance & Landerberg, 2000) and suggested that MG-CFA is one of the most common latent variable approaches. Thus, my dissertation is focused on ME/I in the MG-CFA framework.

In MG-CFA, two types of estimators are developed specifically for ordinal data. The first type is called limited-information estimators because they are based on summary statistics like polychoric correlations (Rhemtulla, Brosseau-Liard & Savalei, 2012); these estimators are commonly used in SEM. The second family of estimators are called full-information estimators; they are often used in IRT. These estimators directly use probit or logit equations to model the relations between latent variables and ordinal indicator, and derive the parameter estimates using case-wise likelihood functions (Bovaird & Koziol, 2012). Past research found that limited-information estimators and the IRT-based full information estimators performed equally well when latent factors are normally distributed (Forero & Maydeu-Olivares,2009; Kim & Yoon, 2011). The full-information estimator was more robust when the latent factors were not normally distributed or the sample sizes were small (Forero & Maydeu-Olivares, 2009; Suh, 2015). However, the full-information estimator has its own limitations. First, when the numbers of latent factors or correlated residuals increase, it may take a long time for the estimator to find the best estimates (Bovaird & Koziol, 2012). Second, it is not available in many SEM software packages (Rhemtulla et al., 2013). For these reasons, I only consider limited-information estimators in the simulation studies.

1.2 Procedures of testing ME/I with MG-CFA

Using MG-CFA, ME/I is usually examined by a series of nested model comparisons through which researchers test whether the measurement parameters such as loadings or thresholds/intercepts of the corresponding indicators are invariant across populations. There are

four levels of ME/I, represented by four invariance models. They are configural, metric, scalar, and strict invariance models, representing least restrictive to most restrictive ME/I. These four models are usually compared in sequence to check whether the factorial patterns, loadings, intercepts, and residual variances are identical across groups (populations) respectively (Kline, 2005). I describe the four models for continuous data first and then show how to extend the models to ordinal data.

In a common factor model, scores on the j_{th} continuous indicator X_j can be expressed as follows:

$$X_j = \tau_{jk} + \sum_{m=1}^{r} \lambda_{jmk} W_m + U_j \,, \qquad (1.2)$$

where r is the number of common factors, p is the number of the manifest measures (j = 1,2……p'), and τ_{jk} is a latent intercept for the j_{th} manifest measure in the k_{th} population. λ_{jmk}, where m =1…r, are the factor pattern parameters (loadings) for the j_{th} manifest variable corresponding to the r common factors in the k_{th} population. W_m is the *m*th factor. U_j is the unique factor for the j_{th} manifest variable. Furthermore, define $E_k(W) = \kappa_k$, $Cov_k(W) = \Phi_k$, $E_k(U) = 0$, $Cov_k(U) = \Theta_k$ (usually assumed to be a p' by p' diagonal matrix, i.e., unique factors are not correlated), $Cov_k(w, \upsilon) = 0$ (null matrix, in CFA, i.e., common factors and unique factors are not correlated).

Based on the above assumptions and definitions, the unconditional moments for manifest variables X can be computed as follows:

$$E_k(X) = \mu_{Xk} = \tau_k + \Lambda_k K_k \,, \qquad Cov_k(X) = \sum\nolimits_{Xk} = \Lambda_k \Phi \Lambda_k' + \Theta_k \qquad (1.3)$$

τ_k is the column vector of τ_{jk}, j=1,2,….*k* (i.e., the column vector of intercepts) in kth population; Λ_k is the matrix of factor loadings for the k_{th} population. For the ME/I tests, one also needs to compute "conditional" moments. The "conditional" means and covariance structure for the manifest measures X given W in the k_{th} population can be represented as:

$$E_k(X|W) = \tau_k + \Lambda_k W \,, \quad Cov_k(X|W) = \Theta_k \qquad (1.4)$$

The parameter matrices in equation 1.4 (i.e., Λ_k, τ_k, Θ_k) are the focus of ME/I tests in the MG-CFA framework. Specifically, given that X has a conditional multivariate normal

distribution (MVN) in each population, ME/I defined in Equation 1.1 holds if

$$\mu_k(X \mid W_t) = \mu(X \mid W_t) = \tau + \Lambda W_t \qquad (1.5)$$

$$\sum\nolimits_{k(X|W_t)} = \sum\nolimits_{(X|W_t)} = \Theta \qquad (1.6)$$

(i.e., Λ_k, τ_k, Θ_k are identical across populations) (Millsap, 2012, p.48, p75-76).

τ, Λ, Θ in equation 1.3 & 1.4 are directly related to the assumptions of configural, metric, scalar, and strict invariance. To test configural invariance, the same factor structure is applied to all groups, but parameters (e.g., loadings, intercepts or thresholds) are allowed to be different across groups. If configural invariance is met, equality constraints could be further imposed on factor loadings to test metric invariance. The null hypothesis for testing metric ME/I model can be represented as:

$$H_0 : \sum\nolimits_{Xk} = \Lambda\Phi_k\Lambda' + \Theta_k, \mu_{XK} = \tau_k + \Lambda K_k \qquad (1.7)$$

If metric invariance is satisfied, equality constraints are added to the corresponding intercepts. The resulting model is the scalar invariance model. The null hypothesis to be tested is

$$H_0 : \sum\nolimits_{Xk} = \Lambda\Phi_k\Lambda' + \Theta_k, \mu_{XK} = \tau + \Lambda K_k \qquad (1.8)$$

Lastly, if the assumption of scalar invariance is tenable, one can further impose equality constraints on the residual variances to establish strict invariance. The null hypothesis to be tested is:

$$H_0 : \sum\nolimits_{Xk} = \Lambda\Phi_k\Lambda' + \Theta, \mu_{XK} = \tau + \Lambda K_k \qquad (1.9)$$

According to the null hypothesis in Equation 1.9 and the ME/I definitions in Equations 1.5 & 1.6, one can tell that, under the MVN assumption, if all the assumptions in a strict invariance model hold, then the ME/I property defined in Equation 1.1 can be established.

Specifically, for the steps described above, global fit indices and chi-square difference tests ($\Delta\chi^2$ tests) are used to determine the plausibility of the invariance models. In the first step when a configural invariance model is tested, global fit indices such as χ^2, CFI, TLI and RMSEA are examined to ensure that the configural model fits the data in the first place. In the second step where the metric invariance model is tested, a $\Delta\chi^2$ is performed between the configural and metric invariance models. If the $\Delta\chi^2$ test is not significant, then the metric invariance model is passed and one can move to step 3. In the third step where the scalar (intercepts) invariance

model is tested against metric invariance model using the $\Delta\chi^2$ test. Similarly, in step 4, strict (residual variance) invariance is tested against the scalar invariance model (Brown, 2006; Kline, 2005).

1.3 Testing ME/I in MG-CFA with ordinal data

Historically, the literature on ME/I in SEM has been mainly focused on continuous indicators (e.g., Meredith, 1993; Widaman & Reise, 1997). Given the popularity of Likert-type indicators in behavioral and social science, researchers have begun to consider the issues raised by original indicators (e.g., Lubke & Muthén, 2004). Sass et al. (2014) examined three methods that researchers commonly use when testing ME/I in SEM with ordinal indicators. These estimators are (1) the continuous maximum likelihood estimator based on normality assumption (ML_{mvn}), (2) the robust continuous maximum likelihood (MLR), and (3) the weight least squared with means and variance adjustment (WLSMV).

1.3.1 Maximum likelihood estimator based on the assumption of normality (ML_{mvn})

Among these three estimators, ML_{mvn} is probably the most commonly used estimator in SEM. It is the default estimator in many SEM software programs for continuous indicators including Mplus (L. K. Muthén & B. O. Muthén,1998-2017) and lavaan in R (Rosseel, 2012). Although ML_{mvn} is based on the normality assumption, some researchers still use it for ordinal indicators if they believe that the continuous and symmetric assumptions are tenable (Li, 2015; Li, 2016). With ML_{mvn}, parameter estimates are obtained through minimizing a fit function (Bollen, 1989):

$$F_{ML} = \ln|\Sigma(\theta)| + trace(S\Sigma^{-1}(\theta)) - \ln|S| - p' \quad (1.10)$$

where θ denotes the vector of model parameters, $\Sigma(\theta)$ is the model implied covariance matrix, S is the sample implied covariance matrix, and p' is the number of observed indicators in the model. With the multivariate normal assumption, the estimates have a standard error that can be computed using the corresponding diagonal elements (variance) in the estimated asymptotic covariance matrix:

$$(\frac{2}{N-1})\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1} \quad (1.11)$$

The test statistic of global model fitness can be then calculated as:

$$T_{ML} = (N-1)F_{ML}\text{ , df = p – q,} \qquad (1.12)$$

where p is the number of the non-redundant elements in sample covariance matrix S, and q is the number of independent parameters in the model. T_{ML} asymptotically follows a χ^2 distribution if the SEM model is correctly specified. T_{ML} is a key statistic for testing ME/I in MG-CFA. Specifically, the T_{ML} statistics of different (nested) invariance models (e.g., $T_{ML\_cofigural}$, an unrestricted model versus T_{ML_metric}, a more restricted model) can be compared through a $\Delta\chi^2$ test:

$$\Delta\chi^2 = T_{ML\_restricted} - T_{ML\_unrestricted}, \qquad (1.13)$$

which follows a chi-squared distribution with df = $df_{restricted\_model}$ - $df_{unrestricted\_model}$.

1.3.2 Robust maximum likelihood estimators (MLR)

A similar but slightly adjusted strategy is to consider the ordinal nature as a kind of non-normality and use the robust maximum likelihood method (MLR) to handle the non-normality (Rhemtulla, et al., 2012). When data are normally distributed, ML_{mvn} estimates are asymptotically efficient; T_{ML} is also asymptotically χ^2 distributed. However, when data are non-normal, the standard error estimates and χ^2 statistic from ML_{mvn} can be biased (see Yuan, Bentler, Zhang, 2005 and Curran, West, & Finch, 1996). Consequently, MLR adjusts the χ^2 statistic and standard errors but point estimates remain the same (Savalei, 2014).

1.3.2.1 A more general form of the fit function and robust standard errors

To illustrate how MLR works, it is helpful to first introduce a more "general" form of the fitting functions in SEM (Jia, 2016; Savalei, 2014). In SEM, researchers assume that the model implied covariance model can be represented as the function of model parameters (i.e., $\sum=\Sigma(\theta)$). The differences between the sample covariance matrix (i.e., S in equation 1.10) and the model implied covariance (i.e., $\Sigma(\theta)$) is the residual variance matrix R. That is

$$S = \Sigma(\theta) + R \qquad (1.14)$$

S and R are p' by p' squared matrices where p' represents the number of observed indicators.

Let s =vector(S) and r = vector(R), which list all of the nonredundant elements in the S and R matrix respectively. vector() is the vectorize function that can turn a matrix into a vector. The relations presented in equation 1.1.4 can be rewritten as:

$$s = \sigma(\theta) + r \qquad (1.15)$$

where $\sigma(\theta)$ is a non-linear function of θ. With these definitions, a more general form of the SEM fit function can be written as

$$F_{genernal} = (s-\sigma(\theta))' W^{-1}(s-\sigma(\theta)) = r' W^{-1} r \quad (1.16)$$

where W is a p by p weight matrix that in practice will be estimated from the data. Let Γ^* be the true asymptotic covariance matrix of s in equation 1.16. If a consistent estimate of Γ^* is used as as the W matrix in 1.16, then asymptotically efficient estimates can be obtained by minimizing $F_{general}$ (Savalei, 2014). The standard errors can be then obtained from the asymptotic covariance matrix:

$$\text{cov}(\sqrt{N}\theta) = (\Delta' W^{-1}\Delta)^{-1} \quad (1.17)$$

where $\Delta = \frac{\partial \sigma(\theta)}{\partial \theta'}|_{\theta}$ (i.e., Δ is the p' by q matrix containing the derivatives of $\sigma(\theta)$ evaluated at θ). The χ^2 statistic can be calculated as $T_{general} = N * F_{genreal}$ (equation A1 in Savalei, 2014).

1.3.2.2 robust chi-square test statistic for continuous non-normality

Let X_i, X_j, X_k, and X_l be the observed variables in the model, and s be the vector of non-redundant elements in the covariance matrix. When data are multivariate normal, the asymptotic covariance matrix of s can be expressed by

$$\Gamma^* = \Gamma_{normal} \quad (1.18)$$

where $\gamma_{normal\_ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$. One can tell that the calculation of each element in the matrix only involves the second moments of the data, which can be easily found in the sample covariance matrix (S) as $\gamma_{normal\_ij,kl} = s_{ik}s_{jl} + s_{il}s_{jk}$. In other words, when data are normal, the consistent estimate of Γ^* can be easily calculated with the elements in S. Asymptotically efficient estimates $\theta_{general,\Gamma_{normal}}$ can then be obtained by using Γ_{normal} as the W matrix in equation 1.16.

When data are non-normal, unfortunately, using Γ_{normal} as the W in equation 1.16 is no longer appropriate. Specifically, when data are non-normal, although $\theta_{general,\Gamma_{normal}}$ is still consistent and asymptotically unbiased, it is not asymptotically efficient. Non-normality affects the standard errors. The test statistic for model fit does not follow a chi-square distribution

asymptotically either.

To solve these problems, Browne (1984) proposed that when data are continuous but non-normal, rather than using the Γ_{normal} as W in equation 1.16, one should use

$$\Gamma^{*} = \Gamma_{ADF} \quad (1.19)$$

where $\gamma_{ADF\_ij,kl} = \sigma_{i\,jkl} - \sigma_{ij}\sigma_{kl}$, as the W matrix. In the sample, $\hat{\gamma}_{ADF\_ij,kl}$ can be calculated as $s_{ijkl} - s_{ij}s_{kl}$. Even though the calculation $\hat{\gamma}_{ADF\_ij,kl}$ involves the fourth moment of the data and makes it far more complicated in comparison to the calculation of Γ_{normal}, it does have its unique advantage over Γ_{normal}. When data are normal, Γ_{ADF} will be equal to Γ_{normal}, and both of them are the consistent estimates of Γ^{*}; in contrast, when data are continuous but non-normal, Γ_{ADF} will still be the consistent estimate of Γ^{*}. Given this property of Γ_{ADF}, researchers can get asymptotically efficient estimates with continuous non-normal data if they are willing to use Γ_{ADF} as the W in equation 1.16. This method is known as the asymptotically distribution free (ADF) estimator (Bollen, 1989).

The ADF method proposed by Browne (1984) has good mathematical properties with infinite sample size. Unfortunately, simulations with finite sample sizes show that the ADF estimates and its χ^2 are only stable with a large sample size (e.g., sample size > = 1000; Curran, West & Finch, 1996). To mitigate the problem, Satorra & Bentler (1994) proposed a different way to correct χ^2 and standard errors (obtained from ML_{mvn}). The corrected χ^2 statistic is calculated as follows:

$$T_{SB} = correctedFactor * T_{ML} = \frac{p-q}{tr(U\Gamma^{*})} T_{ML} \quad (1.20)$$

where $U = \Gamma_{normal}^{-1}[1 - \Delta(\Delta' \Gamma_{normal}^{-1}\Delta)^{-1}\Delta' \Gamma_{normal}^{-1}]$ and $\Gamma^{*} = \Gamma_{ADF}$.

The corrected standard errors of θ are calculated using the following formula:

$$\mathrm{cov}(\sqrt{N}\theta) = (\Delta' \Gamma_{normal}^{-1}\Delta)^{-1}\Delta' \Gamma_{normal}^{-1}\Gamma_{ADF}\Gamma_{normal}^{-1}\Delta(\Delta' \Gamma_{normal}^{-1}\Delta)^{-1} \quad (1.21)$$

where $\Delta = \frac{\partial \sigma(\theta)}{\partial \theta'}|_{\theta}$ (Savalei, 2014). Simulations have shown that in comparison to ADF, Satorra and Bentler's corrections lead to more reliable estimates when the sample size is smaller

than 1000 (e.g., Curran, et al., 1996). Satorra & Bentler (2001) have also developed the rescaled version of the $\Delta\chi^2$ statistic, which can be used to compare the nested ME/I models under non-normality.

Note that although Satorra & Bentler's method performs well on complete data, it cannot accommodate missing data. There are several other versions of robust statistics. Maydeu-Olivares (2017) provided a thorough comparison of the performances of these robust methods. For example, there is a robust method developed by Yuan and Bentler (2000). This method is comparable to Satorra and Bentler's method for complete data but can be applied to incomplete data (see page 23 for details).

1.3.3 Weight least squared estimators based on polychoric correlations

Both ML_{mvn} and MLR treat ordinal data as continuous, so they do not take into account the categorical nature of the data. Some weighted least squares (WLS) estimation methods are designed specifically for ordinal data. In brief, WLS assumes the existence of a normal distributed latent response variate (y_i^*) underlying each ordinal indicator y_i with C categories (e.g., 0,1,2,…..C-1). Every latent response variate y_i^* is categorized by C -1 thresholds ($\tau_{i,C_{i-1}}$, $\tau_{i,C_{i-2}}$,…….. $\tau_{i,1}$) to create the observed ordinal outcomes y_i as:

$$y_i = \begin{matrix} C_i - 1 \\ C_i - 2 \\ \vdots \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} \tau_{i,C_{i-2}} < y_i^* \\ \tau_{i,C_{i-2}} < y_i^* \le \tau_{i,C_{i-1}} \\ \vdots \\ y_i^* \le \tau_{i,1} \end{matrix} \qquad (1.22)$$

The estimation process of WLS involves two or three steps. In the following, I describ the three-step procedure (B. O. Muthén, 1984). First, univariate information from each variable is used to obtain the maximum likelihood estimates of thresholds. These estimates are then treated as fixed in the second step, where researchers use the bivariate information (each pair of observed variables) to calculate the polychoric correlations between each set of paired latent response variates separately. Lastly, estimates are obtained by minimizing the following discrepancy function:

$$F_{WLS} = (s' - \pi(\theta))^T W^{-1} (s' - \pi(\theta)) \qquad (1.23)$$

where s contains the thresholds and polychoric correlation coefficients obtained by the researchers in steps 1 and 2, and $\pi(\theta)$ and θ denote model specifications and model parameters,

respectively. Let Γ^{**} represent the true population asymptotic covariance matrix of S'. A consistent estimate of Γ^{**} should be used as the W matrix in 1.22 (i.e., use Γ_{WLS} as W) (formula 4 in B. O. Muthén, du Toit, & Spisic, 1997). The chi-squared statistic of WLS can be then calculated as:

$$T_{WLS} = (N-1)^* F_{WLS,} \quad df = \quad p^* - q \qquad (1.24)$$

where p* is the number of non-redundant elements in S'.

Note that unlike the WLS estimator used in a general linear model framework, the WLS in SEM is based on summary statistics (e.g., polychoric correlations in S') rather than raw data. Savalie (2014) provided detailed explanations and comparisons between estimators used in regression and SEM frameworks.

WLS provides consistent and asymptotically normally distributed estimates (B. O. Muthén, 1984; B. O. Muthén & Satorra, 1995) when the sample size is large. However, similar to ADF, the performance of WLS is not satisfying with small or medium sample sizes. For example, Flora & Curran (2004) found that when the sample size was lower than 500, WLS was likely to generate improper solutions or convergence problems. The loading estimates and χ^2 from WLS were also substantially biased when sample size was smaller than 500 in a single-group CFA model. In fact, the type I error rate associated with the χ^2 could be substantially inflated in some conditions when the sample size was 1000. These properties severely limit the applicability of WLS in SEM. A solution for this limitation is to invert only the diagonal elements of the weight matrix, rather than the whole W matrix in equation 1.23 (B. O. Muthén, et al., 1997). This approach is named the diagonal weighted least squares estimation method (DWLS). For DWLS, the fitting function is as follows:

$$F_{DWLS} = (S' - \sigma'(\theta))^T (W_D)^{-1} (S' - \sigma'(\theta)) \qquad (1.25)$$

where $W_D = \text{diag}(\Gamma_{WLS})$. Because only part of the weight matrix is inverted, there is a loss of information in estimating the parameters. The loss of the information can be taken into account by applying robust corrections on the chi-square statistics and standard errors (B. O. Muthén, 1998-2004; Savalei, 2014). There are several correction methods that researchers could use. The WLSMV estimator is one of these corrections, and has been found to outperform the others

(DiStefano & Morgan, 2014). WLSMV uses the discrepancy function shown in equation (1.25) to estimate the parameters and corrects the global fit statistic such that the mean and variance of the fit statistic will approximate those of a χ^2 distribution with corresponding degrees of freedom (DiStefano & Morgan, 2014). Specifically, the χ^2 of a WLSMV (T_{DWLS}) is calculated as follows:

$$T_{DWLS} = [d / tr(U'\Gamma^{**})^2]T_{WLS} \quad (1.26)$$

where $U' = (W_D^{-1} - W_D^{-1}\Delta(\Delta' W_D^{-1}\Delta)^{-1}\Delta' W_D^{-1})$, $\Delta' = \frac{\partial \sigma'(\theta)}{\partial \theta}$, and d is the integer closest to $d^* = [tr(U'\Gamma'))^2 / tr((U'\Gamma')^2]$ (i.e., formula 14 – formula 16 in B. O. Muthén, et al., 1997; Li, 2015). The asymptotic covariance matrix of the estimates is calculated as follows:

$$N^{-1}(\Delta' W_D^{-1}\Delta')^{-1}\Delta' W_D^{-1}\Gamma^{**}W_D^{-1}\Delta'(\Delta' W_D^{-1}\Delta')^{-1} \quad (1.27)$$

(i.e., formula 10 in B. O. Muthén, et al., 1997).

WLSMV is the default estimator in SEM software such as Mplus and lavaan when researchers specify their data as ordinal (L. K. Muthén & B. O. Muthén. 1998-2017; Rosseel, 2016). Its robust $\Delta\chi^2$ tests between nested models (e.g., invariance models) can be conducted with the DIFFTEST command with estimator = WLSMV in Mplus (formula 119-121 in B. O. Muthén, 1998-2004; Asparouhov & Muthén, 2006; Asparouhov & Muthén, 2010a).

1.3.4 The relative performance of the estimators for ordinal data

Several studies have been conducted to compare the relative performances of ML_{mvn}, MLR, and WLSMV on ordinal data. In the context of a single group CFA, Rhemtulla et al. (2012) compared the performance of MLR and WLSMV in Mplus. They found that when the number of categories per item is less than five, MLR underestimated the loadings and corresponding standard errors. In comparison, WLSMV yielded more accurate estimates for loadings and standard errors even when the number of categories was seven (especially when thresholds were asymmetric). As for the more general SEM models, Li (2016) found that WLSMV outperformed ML when estimating structural parameters in SEM as well.

In the context of ME/I testing using MG-CFA, Sass et al. (2014) found that $\Delta\chi^2$ tests from ML_{mvn} and MLR slightly outperformed WLSMV in detecting non-invariant thresholds with small sample sizes; in contrast, the $\Delta\chi^2$ test of WLSMV (DIFFTEST function in Mplus)

performed better on detecting non-invariant loadings. However, the results obtained from Li (2016), Rhemtulla et al. (2012), and Sass et al. (2014) are all based the assumption that researchers have ordinal "complete" data. I discuss the ordinal missing data issue in section 1.5 and its potential influences on ME/I tests in section 1.6.

1.4 Specification search – identify the non-invariant items

In sections 1.2 and 1.3, I reviewed the general procedure for ME/I tests and the variety of estimators that can be used when encountering ordinal data. In this section, I discuss some post hoc modifications that researchers might adopt after a specific level of ME/I failed. In brief, the ME/I tests described above can be thought of as omnibus tests, similar to F tests in ANOVA. For example,when the $\Delta\chi^2$ test rejects the assumption of metric invariance, it indicates that at least one equality constraint on loading should be released, but not which constraint(s) is implausible. Therefore, other statistics (e.g., modification indices, MFI) need to be used to identify (i.e., screen out) the non-invariant items. Once the non-variant items are identified, one may release the corresponding constraints (or discard the non-invariant items if it can be justified; Millsap & Kwok, 2004). This model modification process is called a specification search (Millsap, 2012, p.108). Past research has demonstrated the importance of specification search (i.e., it correctly finds the invariant and non-invariant items). Specifically, failing to identify the non-invariant items and release the misspecified constraints can cause biased cross-populations (groups) comparisons (e.g., Guenole & Brown, 2014; Chen, 2008). On the other hand, if the researchers correctly impose the equality constraints, it can increase the power of group comparison tests (Xu & Green, 2015).

Specification search after ME/I tests in MG-CFA can be done either backwardly or forwardly (see Jung & Yoon, 2016; Yoon & Kim, 2014). For example, if a metric invariance model is rejected, one can start from the configural invariance model and gradually impose equality constraints. This is called a forward specification search. Alternatively, one may start from the metric invariance model and gradually release the equality constraints. This is called a backward procedure.

Researchers have shown that the backward approach is effective when the proportion of non-invariant items in the model is low (Yoon & Millsap, 2007). Among the different backward approaches, specification search based on the largest modification index is most advantageous (Jung & Yoon, 2016; Yoon & Kim, 2014). Thus I briefly explain this backward approach in

section 1.4.1. In section 1.4.2, I introduce a recently proposed forward specification search based on confidence intervals, which showed comparable performance to the backward search with modification indices in literature.

1.4.1 Backward specification search based on the largest modification indices (backward MFI method).

The backward MFI method uses the largest modification index to identify non-invariant items. A modification index (MFI) is a univariate Lagrange multiplier that is expressed as the chi-squared statistic with df =1 (Bollen, 1989, page 299). MFI captures the amount by which the model χ^2 statistic will decrease after a model constraint is released (Kline, 2005, page 217). The backward MFI method first releases the equality constraint with the largest MFI that is greater than a pre-set cutoff value (e.g., > 3.841 at α=.05). After releasing the constraint, researchers can reevaluate the MFIs for other constraints to see whether there is another constraint to release. The researchers continue doing this until no MFI on the equality constraints is larger than the cutoff (Yoon & Millsap, 2007).

Even though it is still prone to the type I error rate, simulations found that the backward MFI method outperformed other specification search methods, such as the factor-ratio test or the non-sequential specification search method (Jun & Yoon, 2016; Yoon & Kim, 2014). Besides, given the fact that backward MFI starts from the most restrictive model (e.g., the full metric invariance model), researchers need not worry about the anchor item problem when conducting the specification search (Yoon & Millsap, 2007). This is one of the advantages of the backward approach to the forward specification search methods.

Some aspects of the backward MFI method are worth mentioning. First, given that MFI is based on the likelihood function of the model and closely related to the χ^2 statistic, if the estimation method changes, the value of MFI will also change. For example, in Mplus, if researchers change the estimator from ML_{mvn} to MLR, the MFI will be adjusted as the χ^2 statistic (L. K. Muthén, 2011). Second, researchers have found that the performance of MFI is affected by sample size. As sample size decreases, the power of the MFI also decreases (Chou & Bentler, 1990).

1.4.2 A new proposed, forward specification search method based on confidence intervals. (forward CI method)

Although the backward MFI method has been widely used in empirical studies (Yoon & Kim, 2014), researchers are still developing new specification methods. Recently, Jung and Yoon (2016) proposed a forward specification search method based on confidence intervals (the forward CI method). This method does not use MFI. Instead, it relies on the confidence intervals of differences between corresponding parameters. Specifically, with the forward CI method, researchers first define new parameters to describe the difference between corresponding parameters (e.g., newParameter = loading_1 in group 1 minus loading_1 in group 2). After that, they can use the "model constraints" command in Mplus to calculate the confidence intervals of these newly-defined parameters. The "model constraints" command in Mplus uses the "delta method" to estimate the standard error of the newly-defined parameter (L. K. Muthén, 2010). Assume $g(\theta)$ is the function that researchers use to create a new parameter based on the original model parameters θ. The delta method uses the first order term in the Taylor series expansions of $g(\theta)$ to approximate $g(\theta)$ (Casella & Berger 2002). Using the approximate term rather than the original $g(\theta)$ could make the estimations of mean and variance of the newly defined parameter easier. According to Jung & Kim (2016), if the $(1-\alpha)$% CI of a new defined parameter covers zero at a certain alpha level (e.g., α=.05 or .01), it means the parameters used to create the new parameter are not invariant (i.e., they are identified as non-invariant items).

The method proposed by Jung & Yoon (2016) is a "forward" search method in the sense that it always starts from the less restricted invariance model. For example, to screen out non-invariant loadings, researchers will fit the configural invariance model with "model constraints" comments based on corresponding loadings. Similarly, to detect the non-invariant intercepts/thresholds, they will fit the metric invariance with the "model constraints" comments based on corresponding intercepts/thresholds across groups. By starting with a less restricted model, Jung & Yoon's method is less prone to misspecification. However, it requires at least one anchor item before searching; otherwise, the model might be unidentified.

Jung & Yoon (2016) found that the performances of their forward CI method were comparable and generally outperformed the backward MFI method from the perspectives of perfect recovery rate and model level type I error rate. However, an assumption made in Jung & Yoon (2016) is that the anchor items are correctly specified. This is an unnecessary assumption

for the backward MFI method.

1.5 Missing data in structural equation modeling

In sections 1.3 and 1.4, I reviewed the studies about ME/I tests and specification searches in MG-CFA. However, all of the studies considered only complete data, which is a limitation because missing data is a common problem in SEM (Marsh, 1998).

In SEM literature, the relative performances of missing data methods (MDTs) under different kinds missing data mechanisms have been addressed (e.g., Enders, 2001; Enders 2001; Enders & Bandalos, 2001; Marsh, 1998; Savalei & Bentler, 2005; Teman, 2012; Wu, Jia & Enders, 2015). According to the classification in Little & Rubin (2002), there are three missing data mechanisms missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR indicates that the presence of the missing data is independent of both observed and missing data; MAR indicates that the presence of missing data is dependent on observed data but independent of missing data; otherwise, the data are missing not at random (MNAR).

There are traditional and modern methods to deal with missing data (Enders, 2010). Traditional methods include listwise deletion, pairwise deletion (PD), and single imputation methods. Modern missing data methods include full information maximum likelihood (FIML) and multiple imputations (MI). Historically, traditional methods like listwise and pairwise deletion methods were the most common MDTs in SEM (Marsh, 1998). However, studies revealed that the estimates obtained from modern MDTs such as $FIML_{mvn}$ and MI are in general more accurate and efficient (e.g., Enders & Bandalos, 2001; Teman, 2012; Wu et al., 2015) because they can use all the available information in the data and account for the uncertainty caused by missing data when estimating standard errors. Simulations have consistently found that modern MDTs generally outperformed the traditional MDTs when data are MAR or MNAR (Enders, 2010). The performance of FIML and MI in the SEM literature is discussed in the following two sections.

1.5.1 Full information likelihood method based on the assumption of normality ($FIML_{mvn}$)

FIML maximizes the sum of the log of "casewise" likelihood functions to obtain estimates. Based on the assumption that data follow a multivariate normal distribution (MVN), the log likelihood of case i can be defined as

$$\log L_i = K_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(X_i - \mu_i)'\sum_{i=1}^{N}{}^{-1}(X_i - \mu_i) \quad (1.28)$$

, where K_i is a constant based on n_i. On the other hand, the log likelihood of the whole sample will then be

$$l(\theta) = \log L(\mu, \Sigma) = \sum_{i=1}^{N} \log L_i \quad (1.29)$$

By maximizing equation 1.29, parameter estimates are obtained ($FIML_{mvn}$, Arbuckle, 1996, p248, Yuan & Bentler, 2000, p.167). The test statistic of $FIML_{mvn}$ can be then calculated as

$$T_{FIML_{mvn}} = -2(l(\theta) - l(\beta)) \quad (1.30)$$

where the $l(\theta)$ is the maximized log likelihood under the tested model and $l(\beta)$ is the corresponding maximized log likelihood under the saturated model. β are the estimates of the saturated model parameter β (Yuan & Bentler, 2000).

When data are complete, the estimates obtained from equation 1.29 are identical to the estimates obtained from equation 1.10 (Savalei, 2008). However, when data are incomplete, using equation 1.29 allows researchers to avoid using listwise or pairwise deletion methods to calculate the sample covariance matrix S in equation 1.1, given that participants with different missing data patterns are allowed to have their own likelihood functions in equation 1.29.

Enders & Bandalos (2001) found that $FIML_{mvn}$ generated accurate estimates of loadings and had a good control on the type I error rate under both MCAR and MAR conditions. In contrast, the χ^2 tests obtained from pairwise deletion and similar response pattern imputation methods led to highly-inflated Type I error rates, even under MCAR. For the listwise deletion method, although the type I error rate of χ^2 was mostly acceptable under MCAR, it produced substantially biased loading estimates under MAR. Marsh (1998) also investigated the point estimates and χ^2 obtained from pairwise deletion in MCAR. He found that even though the loading estimates were generally accurate, the χ^2 statistic was substantially biased.

Note that although traditional deletion methods have multiple disadvantages in comparison to the modern MDTs, they are still the default MDTs for the DWLS estimators in multiple SEM software packages. For example, listwise and pairwise deletion methods are default MDTs for WLSMV in the lavaan package in R and Mplus, respectively (e.g., Rosseel, 2017). Furthermore,

researchers sometimes have to use these deletion methods, because there is no better way to calculate the $\Delta\chi^2$ statistic and MFI for ME/I tests with modern MDT such as multiple imputations (e.g., Zhou et al., 2016). This problem was further illustrated in section 1.6.

1.5.2 Robust full information likelihood method (Robust FIML)

Simulations indicated that $FIML_{mvn}$ performs well in SEM when data were multivariate normally distributed (e.g., Arbuckle, 1996; Enders & Bandalos, 2001). However, researchers also found that when data are continuous and non-normally distributed, both standard errors and χ^2 statistics obtained from $FIML_{mvn}$ could be biased (Enders, 2001). Fortunately, similar to corrections developed for χ^2 statistic and standard errors with complete data, robust $FIML_{mvn}$ has also been developed to correct for the influences of continuous non-normality with incomplete data. Yuan and Bentler (2000) extended Satorra and Bentler's work to the scenario of incomplete data. In the current study, I use the term robust FIML for their method. With robust FIML, the standard errors are calculated using a sandwich-type covariance matrix:

$$\text{cov}(\sqrt{N}\theta) = A^{-1}BA^{-1} \qquad (1.31)$$

where A = $-\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 l_i(\theta)}{\partial\theta\partial\theta'}$, B = $-\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\theta)}{\partial\theta}\frac{\partial l_i(\theta)}{\partial\theta'}$. The derivatives in A and B are both evaluated at θ; l_i is the log likelihood of case *i* in the structured model (formula 9a and 9b in Yaun & Bentler, 2000). The χ^2 statistic from robust FIML is calculated as follows:

$$T_{robsut\_FIML} = \frac{d}{tr(\Omega_\beta U)} T_{FIML_{mvn}} \quad (1.32)$$

Ω_β in equation 1.32 is an estimate of the asymptotical covariance matrix for the saturated model. Specifically,

$$\Omega_\beta = \text{cov}(\sqrt{N}\beta) = A_\beta^{-1} B_\beta A_\beta^{-1} \qquad (1.33)$$

where A_β is the observed information matrix assuming normality; B_β is the covariance matrix of the first derivatives of $l_i(\beta)$; l_i is the normal theory log-likelihood that researchers use to obtain the estimates of the saturated model. Both Ω_β and A_β are evaluated at the tested model.

U can be calculated as:

$$U = A_\beta - A_\beta \Delta (\Delta' A_\beta \Delta)^{-1} \Delta' A_\beta \quad (1.34)$$

where $\Delta = \frac{\partial \beta(\theta)}{\partial \theta'}|_{\theta=\theta}$ is the matrix of model derivatives evaluated at the FIML estimates θ (Savalei & Bentler, 2005; Savalei & Folk, 2014). Enders (2001) showed that robust FIML was able to mitigate the influence caused by non-normal data. However, there are conditions under which robust FIML was not effective (Savalei & Folk, 2014).

1.5.3 Multiple imputations (MI)

MI is another "modern" missing data method which will lead to accurate point estimates and standard errors under MCAR and MAR (Enders, 2010; Schafer & Graham, 2002). Unlike traditional single imputation methods, MI generates multiple independent imputed data sets to take into account the uncertainty caused by missing data into estimation (Enders, 2010). MI involves three phases: (1) generating multiple imputed datasets, (2) conducting analysis for each imputed data set, and (3) pooling analysis results across imputed data sets. These phases are explained in the following sections.

1.5.3.1 The imputation phase

With continuous data, MI (MI_{mvn}) is often conducted by using the data augumentation algorithm (Enders, 2010). The algorithm starts with some initial values on imputation parameters (e.g., elements in the mean vector and covariance matrix) and then iterates between an imputation step (I-step) and a posterior step (P-step) (Schafer, 1997). At the P-step, a random set of imputation parameters $\theta^{(j)}$ is drawn from their posterior distribution. In the I-step, missing data (i.e., Y_{miss}) are predicted by Y_{obs} based on $\theta^{(j)}$, which is equivalent to drawing random values from the predictive probability distribution of Y_{miss} (e.g., multivariate normal distribution).

Specifically, the P-step and I-step at the j^{th} iteration can be considered as drawing $\theta^{(j)}$ from $p(\theta | y_{obs}, y_{miss}^{j-1})$ and $y_{miss}^{(j)}$ from $p(y_{miss} | y_{obs}, \theta^{(j)})$, respectively. For example, given starting values on $\theta^{(0)}$, the data argumentation algorithm could start with the I-step by generating predicted values of Y_{miss}. These predicted values and Y_{obs} are then carried to the P step in the first iteration to re-define (or update) the posterior distributions of imputation parameters from which $\theta^{(1)}$ is drawn. This cycle will repeat itself until the posterior distribution of imputed parameters is stable (i.e., the model has converged). When the posterior distribution

stabilizes, the imputed data sets can be saved. To avoid autocorrelation between adjacent iterations, rather than directly saving the results of first m iterations, it is recommended to save imputed data at every k^{th} iteration until a desirable number of imputed datasets is reached.

1.5.3.2 The analysis phase

In the second phase of MI, the target analysis (e.g., regression) is applied to each of the imputed data set separately. For example, suppose there are 5 imputed data sets; researchers fit the regression model to these 5 imputed data sets separately to get 5 regression coefficients, 5 intercepts, and their corresponding standard errors.

1.5.3.3 The pooling phase

The third phase of MI is to pool these parameter estimates and standard errors across imputations. Rubin (1987) proposed rules to pool the point estimates across imputed data sets. In brief, the point estimates (e.g., the regression coefficient or the intercept of a simple regression model) can be calculated simply by taking the average of point estimates across imputations (i.e., $\overline{\theta} = \frac{\sum_{m=1}^{M}\theta_m}{\mathrm{M}}$, where θ_m represents the parameter estimates in m^{th} imputation). Pooling standard errors across imputations is more complicated. The average of standard errors in each data set only represents the sampling fluctuation had the data been complete ($V_w = \frac{\sum_{m=1}^{M} SE_m^2}{M}$). To effectively capture the additional variability caused by missing data, another source of sampling variance, V_b, (variance between imputations) should be included. The total sampling variance (V_t) of the estimates is the sum of the two variances plus an adjusted term that reflects the fact that the variance in estimates will decrease as the number of imputations is increased, $V_t = V_w + V_b + \frac{V_b}{M}$. The final estimates of standard error after a finite number of imputations are then the squared root of V_t.

By using the pooled point estimate and standard error, one can conduct hypothesis testing of the parameter ($\mathrm{H}_0 : \theta = \theta_0$) using a t statistic, where $t = \frac{\overline{\theta} - \theta_0}{\sqrt{V_T}}$. df = $v = (m-1)(1 + \frac{V_w}{V_B + V_B / m})$. It is noticeable that when the number of imputations (m) increases,

the $df -> \infty$ and t -> Z; then the t-test will approximate the Wald test obtained from $FIML_{mvn}$ (Enders, 2010, p.231).

1.5.4 Multiple imputation methods for ordinal missing data in SEM

Although the imputation phase of MI described above is based on the assumption of multivariate normality, it can be easily adopted for ordinal data. There are several ways to adjust the imputation for ordinal missing data. These methods include:

(1) **Impute by treating the ordinal data as if they were continuous.** Researchers can determine whether to round off the continuous imputed values depending on the follow-up analysis.

(2) **Impute missing data using categorical data approach such as discriminate analysis or logistic regression**. This type of imputation method will generate ordinal imputed data, according to the posterior probabilities of categories within an item (Burren & Goothuis-Oudshoorn, 2010).

(3) **The latent variable approach** (Wu, Jia & Enders, 2015). This approach imputes missing data at the latent variables level first. The latent variables are assumed to follow a multivariate normal distribution. The imputation process is similar to what is described in section 1.5.3.1. Specifically, it is assumed that latent response variates y^* underlying ordinal indicators jointly follow a multivariate normal distribution (i.e., $y^* \sim MVN(\mu, \Theta)$), where the diagonal elements of Θ are fixed at 1 (to set the scale) and other elements are freely estimated. The latent variable approach first imputes the data at the y^* level. These continuous data will then be discretized with equations similar to equation 1.22, where threshold parameters are drawn from their posterior distribution (Asparouhov & Muthén, 2010b).

Past research has shown that the latent variable approach works well for ordinal missing data when the analysis is done using WLSMV (Asparouhov, & Muthén, 2010c; Teman, 2012). Asparouhov & Muthén, (2010c) showed that it can help researchers obtain more accurate point estimates in a latent growth curve model than WLSMV plus pairwise deletion. Teman (2012) also found that the latent variable approach worked well with WLSMV for CFA models with ordinal missing data. Specifically, Teman found that it outperformed $FIML_{mvn}$ or WLSMV combined with listwise deletions in terms of loading estimates in CFA models with five-point

ordinal indicators. In the current study, I will refer to the combination of WLSMV and the latent variable imputation method as WLSMV_MI.

According to literature, WLSMV_MI might be one of the most effective methods for handling ordinal missing in SEM thus far. Besides the good performance of this method found in Asparouhov & Muthén (2010c) and Teman (2012), the latent variable approach will generate imputed data that follow the original ordinal metric. Thus, estimators based on polychoric correlations like WLSMV can be used to analyze the imputed data. In contrast, the normal imputation approach will only generate continuous data, so WLS type estimators cannot be used in the data analyses following imputation.

Note that so far, MI is still the only modern MDT that can be used with WLS esimators. FIML (or robust FIML), is not available when WLS estimators are used. This is due to the limited information nature of WLS estimators. As mentioned in section 1.3.3, the estimation process of WLS based estimators involves multiple stages (three or two). During the first two stages, where the thresholds and polychoric correlations are estimated, only "univariate" or "bivariate" information is used, thus it is impossible to use "full" information likelihood (or the casewise likelihood function) in these stages. If there are missing data, then they will be listwise or pairwise deleted by default. In Mplus, pairwise deletion (PD) is the default for WLSMV.

1.6 Issues caused by ordinal missing data on ME/I tests and specification searches.

Based on my review of the literature on ordinal data, missing data, ME/I tests, and specification search methods in SEM, the major findings in the literature are summarized below.

(1) Treating ordinal data as continuous can cause biased point estimates. Thus, DWLS estimators such as WLSMV are a more appropriate choice with complete data (Li, 2016, Rhemtulla et al., 2013).

(2) When missing data are present, it is better to use the modern MDT such as FIML or MI to handle missing data. Traditional MDTs such as listwise deletion or pairwise deletion can cause substantially biased results (e.g., Enders, 2001; Enders & Bandalos, 2001).

(3) Given that FIML cannot work with DWLS estimators, MI is a reasonable choice for researchers. Methodologists found that WLSMV_MI is an effective way to handle ordinal missing data in SEM (Asparouhov & Muthén, 2010c; Teman, 2012). However, WLSMV_MI has limitations in ME/I tests and specification search procedures.

Specifically, neither $\Delta\chi^2$ nor modification indices are available with WLSMV_MI (Liu, et al., 2017; B. O. Muthén, 2017). These limitations are further explained below

1.6.1 Potential effects of ordinal missing data on ME/I tests

The first problem with applying the WLSMV_MI method in ME/I tests with ordinal missing data is that the $\Delta\chi^2$ test is not available. As mentioned before, an important phase in MI is pooling the statistics across imputed data sets. Unfortunately, there is thus far no agreed upon way to appropriately pool the χ^2 statistics across imputed data sets with WLS estimators (or more precisely, any non ML_{mvn} estimators after MI, Enders, 2010; Liu, et al. 2016). Mplus will only provide an average χ^2 across imputed data sets after WLSMV_MI. Teman (2012) found that this average χ^2 (with an average degree of freedom) should not be directly used for χ^2 tests, because it severely inflated type I error rate. Given this limitation, if researchers have to use the $\Delta\chi^2$ tests, they need to use the ill-advised traditional MDTs, such as the pairwise deletion method for WLSMV. In this study, I call this combination WLSMV_PD. In theory, WLSMV_PD should only be used when data are MCAR or a special type of MAR where the missingness is only determined by independent covariates X in the model (Asparouhov & Muthén, 2010d). Asparouhov & Muthén denoted this kind of special case of MAR as MARX. It is important to point out that if any indicator of a CFA model has MAR data, then MARX does not hold and WLSMV_PD will lead to biased point estimates. Regardless of the problem, some researchers may still prefer WLSMV_PD for ME/I tests if they believe that correctly modeling the ordinal nature of data is more important than the potential information lost and bias in parameter estimates and χ^2 test statistic caused by using pairwise deletion (e.g., Zhou et al., 2016) even when the data are not MCAR or MARX.

On the other hand, some researchers might have quite different opinions on how to handle the ordinal missing data problem in their ME/I tests. For example, Fokkemma et al. (2013) explicitly mentioned that in comparison to using DWLS estimators, which by default might force them to use those ill-advised traditional MDTs, they prefer to treat ordinal data as continuous (with ML_{mvn} or MLR) so that they can use the modern MDT such as $FIML_{mvn}$ and robust FIML for missing data. In other words, researchers such as Fokemma et al. (2013) believe that being able to use a better missing data method is more important for ME/I tests than correctly modeling

the ordinal nature of the data.

1.6.2 Potential effects of ordinal missing data on specification searches

The same problem applies to the backward MFI specification search. Researchers are faced with two options, each of which has its pros and cons. Given the lack of an appropriate way to pool MFI after MI (B. O. Muthén, 2017), researchers who prefer to use WLS estimators might use PD even though PD is known to be problematic; otherwise, they may choose to use FIML (or robust FIML), for which the MFI can be directly computed, by treating ordinal data as continuous.

A possible solution for the dilemma in backward MFI searches is to use the forward CI search method with WLSMV_MI because the CIs be obtained with WLSMV_MI. In Mplus, if one applies the forward CI method with WLSMV_MI, Mplus will apply the delta method to every imputed data set to obtain standard errors and then pool them together with the formula described in Asparouhov & Muthén (2008) (Asparouhov, 2017). This method is theoretically appealing given that researchers can use a modern missing data technique and account for the ordinal nature of the data at the same time. However, this approach hasn't been thoroughly examined in previous research. The relative advantages and disadvantages of different ME/I testing and specification search methods are summarized in Table 1.

Table 1. *Pros and cons for using different methods to test measurement invariance and conduct specification search with ordinal missing data*

	Using modern missing data methods?	Treating data as ordinal?	Note:
ME/I testing			
FIML	Yes	No	$\Delta\chi^2$ tests, point estimates and their SE could be affected due to treating ordinal data as continuous
Robust FIML	Yes	No	Same as above, but have robust corrections on $\Delta\chi^2$ tests and point estimates' SE
WLSMV_PD	No	Yes	$\Delta\chi^2$ tests, point estimates and their SE could be affected due to using pairwise deletion
WLSMV_MI	Yes	Yes	$\Delta\chi^2$ test is not available
Search			
FIML (MFI)	Yes	No	A backward search method based on modification indices (MFI). Its MFI could be affected due to treating ordinal data as continuous
Robust FIML (MFI)	Yes	No	Same as above, but have robust corrections on MFI
WLSMV_PD (MFI)	No	Yes	A backward search method based on modification indices (MFI). Its MFI could be affected due to pairwise deletions
WLSMV_MI (CI)	Yes	Yes	Modification indices are not available ;while it can be used with a recently proposed forward search method based on confidence interval (Jung & Yoon, 2016). It has the potential to ordinal and missing data problem simultaneously in theory; while this idea has never been tested

Chapter 2─Research Questions

In chapter 1, I reviewed the literature related to the ordinal missing data problem in ME/I tests and specification searches. In this chapter, I briefly summarize the limitations of the previous research which bring up four research questions for my dissertation.

2.1. Limitations of Previous Research

Even though past research suggested that WLSMV_MI can produce accurate parameter estimates in SEM with ordinal missing data, this approach is limited in ME/I testing because there is no appropriate way to pool χ^2 and MFI statistics across imputed data sets with this approach. Consequently, "suboptimal" strategies may have to be used such as: (1) misspecify ordinal data as continuous so that $FIML_{mvn}$ or robust FIML can be used; (2) stay with ordinal estimators such as WLSMV, even though it will require the use of the not-recommended, deletion-based MDTs.

According to the literature, these suboptimal approaches both have their own weaknesses, given that treating ordinal data as continuous and using the traditional MDTs are both known to be problematic. However, few studies have compared the performances of these "suboptimal" methods in the context of ME/I testing. To fill in this gap in the literature, I conduct two simulation studies to examine the performances of these methods. In addition, given that the forward CI specification search method proposed by Jung & Yoon (2016) has the potential to work with modern MDT without treating ordinal data as continuous, I also consider the method in my dissertation. The results of these studies could provide useful guidance to applied researchers in selecting more appropriate methods for ME/I tests and specification searches with the presence of ordinal missing data.

2.2 Research questions

Specifically, my dissertation aimed to address the following research questions:

Question 1: How do the different strategies ($FIML_{mvn}$, robust FIML, and WLSMV_PD) perform in terms of $\Delta\chi^2$ tests for measurement invariance testing with ordinal missing data?

Hypothesis 1: previous studies on ML_{mvn} with continuous data have shown that using pairwise deletion can result in type I error rates inflation for χ^2 tests (e.g., Marsh, 1998). Thus, I expected that that the performance of $\Delta\chi^2$ test from WLSMV_PD will only be acceptable when data are

complete. Its type I error rates will be inflated as missing data present. In contrast, FIML and robust FIML will be more robust to the missing data. Robust FIML might slightly outperform $FIML_{mvn}$ given it corrects for non-normality of ordinal data.

Question 2: How do the different strategies (FIML, robust FIML, and WLSMV_PD perform in producing accurate parameter estimates and standard error estimates?

Hypothesis 2: Previous studies have found that WLSMV_PD generated accurate factor loading estimates even with ordinal missing data. In contrast, $FIML_{mvn}$ and robust FIML treat the ordinal data as continuous data. Thus, I expected that the WLSMV_PD will produce the most accurate loading estimates than $FIML_{mvn}$ or robust FIML. As for standard errors estimates,.

Question 3: How do the different strategies (FIML, robust FIML and WLSMV_PD) perform in backward MFI search procedures?

Hypothesis 3: Given that previous research found that pairwise deletion inflated the χ^2 tests in SEM, I expected that similar result will be found in specification search that based on MFI. That is, the type I error rates of the specification search from WLSMV_PD will be substantively inflated as missing data rate increases. In contrast, the type I error rates from the two FIML methods will be more robust to missing data and both will outperform WLSMV_PD in terms of perfect recovery rates (i.e., correctly locate all non-invaraint parameters without mis-identifying any invariant parameters) and controlling for type I error rates.

Question 4: How does the forward specification search with confidence intervals perform using WLSMV_MI in comparison to the backward search methods using the three strategies in question 3?

Hypothesis 4: Given that WLSMV_MI is the only specification search method in the current studies that can use the modern missing data techniques with ordinal estimators, I expected that it will outperform all other three strategies (FIML, robust FIML and WLSMV_PD).

Chapter 3─Simulation Studies

In this chapter, I present two simulation studies to address the four research questions mentioned above. The first study was designed to address the first two research questions related to global ME/I tests; the second study was designed to address the two research questions about specification search. The research designs for the two studies were described in sections 3.1 and 3.2.

Note that I use the term "invariant conditions" to indicate the between-replication conditions where the items are all invariant. I used the term "non-invariant conditions" to represent the between-replication conditions where some of the items are non-invariant. Non-invariant conditions are further categorized as either loading non-invariant conditions or threshold non-invariant conditions, depending on the type of non-invariant parameters.

3.1 Study 1

The population model for this study was specified based on Sass et al. (2014) and presented in Figure 1. Specifically, it is a two-group, single factor CFA model with ten indicators per group. The indicators are 5-point ordinal variables. An auxiliary variable, was used to generate missing data, was created for each group that correlates with the latent factor. Group A is used as the reference group, where the data are complete and the parameters are fixed across all conditions. Group B is the focus group, where the loadings or thresholds in items 8, 9 and 10 were varied in the non-invariant conditions. Specifically, factor loadings of all items in group A and the first seven items in group B were always fixed at 0.6 across all conditions; their thresholds are fixed at -1.3, -0.47, 0.47, and 1.3 in symmetric threshold conditions and fixed at -0.253, 0.385, 0.842, and 1.282 in asymmetric threshold conditions. The loadings or thresholds for items 8 – 10 in group B were subtracted by certain values to create non-invariance. Note that non-invariance occurred in either loadings or thresholds but not both. In addition, for conditions with missing data, missing data were imposed on items 8 – 10 only.

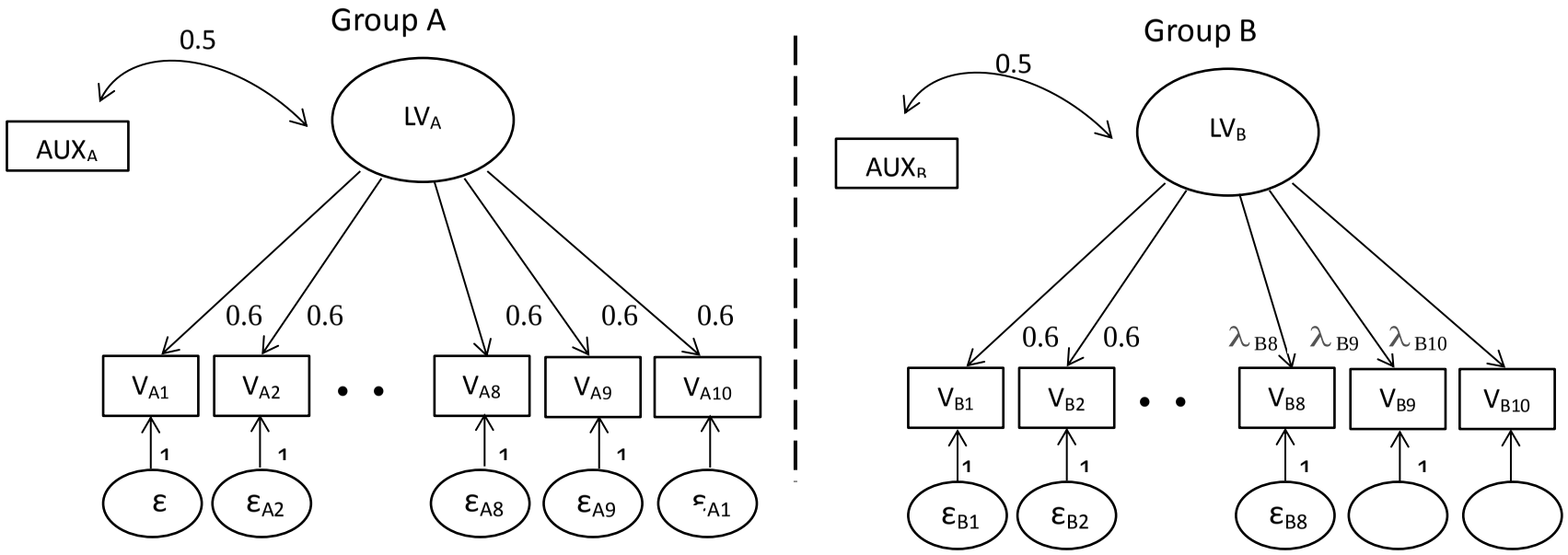

Figure 1. The population model for study 1

Note: AUX_A and AUX_B represent the auxiliary variables for groups A and B. LV_A and LV_B represent the latent variables for groups A and B. LV_A(LV_B) and AUX_A(AUX_B) jointly follow a $MVN(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix})$. V_{A1}–V_{A10} and V_{B1}–V_{B10} are the ten ordinal indicators for each group.

The design factors in the study 1 are summarized as follows:

1) Sample size (150 per group, 250 per group, 500 per group)
2) Locations of non-invariance (invariance, non-invariance in loadings, or non-invariance in thresholds)
3) Amount of non-invariance (0.2, 0.3, 0.4, 0.5). In non-invariant conditions, the new parameter values of items 8–10 in group B will be the original population values subtracted by these values.

The details of the non-invariant parameters in non-invariant conditions are presented in Table 2.

Table 2. *Model parameters of different amounts of non-invariance in study 1*

Parameter	Group A	Group B				
		Amount of non-invariance = 0	0.2	0.3	0.4	0.5
Loading in items 1–7	.6	.6	.6	.6	.6	.6
Loading in items 8–10	.6	.6	**.4**	**.3**	**.2**	**.1**
Symmetric						
Thresholds in items 1–7	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)
Thresholds in items 8–10	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(**-1.5, -0.67, 0.27, 1.1**)	(**-1.6, -0.77, 0.17, 1.0**)	**(-1.7, -0.87, 0.07, 0.9)**	**(-1.8, -0.97, -0.03, 0.8)**
Asymmetric						
Thresholds in items 1–7	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)
Thresholds in items 8–10	(-0.253, 0.385, 0.842, 1.282)	(-0.253, 0.385, 0.842, 1.282)	**(-0.453, 0.185, 0.642, 1.182)**	**(-0.553, 0.085, 0.542, 0.982)**	**(-0.653,- 0.015, 0.442, 0.882)**	**(-0.753,- 0.115, 0.342, 0.782)**

Note: Locations of non-invariance are set to either loadings or thresholds in items 8–10 in group B. Non-invariance occurs on either loadings or thresholds.

4) Distributions of thresholds (symmetric = -1.3, -0.47, 0.47, 1.3; asymmetric = -0.253, 0.385, 0.842, 1.282)

5). Missing data proportions (0%, 30%, 50%)

6). Estimation and missing data methods ($FIML_{mvn}$, robust $FIML_{mvn}$, and WLSMV_PD)

Factors 1 – 4 are between-replication factors, whereas factors 5 and 6 are within-replication factors. The levels of the factors were explained in more detail below. For each combination of the between-replication factors, I generated 500 data sets. The general process of study 1 could be broken into five steps: (1) Generating 500 continuous datasets based on the population values of loadings in a condition. The mean of the latent factors is set to 0 and the variances of the latent factors are set to 1; the residual variances are set to 1 – loadings^2. (More details can be seen in section 3.1.1); (2) categorizing continuous data in each of the datasets based on selected thresholds in that condition. This step generated complete ordinal data; (3) imposing missing data on the last three items (items 8–10) in group B based on the auxiliary variable (described in detail later); (4) sending the complete and missing data sets to Mplus for ME/I tests and the results are recorded; (5) R is used to summarize the results.

In the above process, steps 1 and 2 generate conditions related to the between replication-factors (i.e., sample size, locations of non-invariance, the amount of non-invariance, and distributions of thresholds), and steps 3 to 5 generate conditions related to within-replication factors.

3.1.1 Complete data generation and between replication conditions

I assumed that there was a continuous and normal distributed latent variable underlying each ordinal indicator. Thus, the data were first generated at the latent continuous variable level. Specifically, bivariate normal distributed data are generated using the MASS package within R 3.3.1 (R core team, 2016). For each group, I first used the "mvrnorm" function to generate standardized bivariate normal distributed data X, where the covariance is set to 0.5 (i.e., $x = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$), X_1 is the later factor and X_2 the continuous auxiliary variable that will later be used to generate missing data. The latent factor was multiplied with the population loadings to create the continuous latent response variates underlying each ordinal indicator. After that, for each latent response variate, a normally distributed error term

[$\varepsilon \sim N(\mu = 0, \sigma^2 = 1 - loading^2)$] was added to it. For example, in non-invariant conditions, given that the loadings are always fixed at 0.6, $\varepsilon \sim N(\mu = 0, \sigma^2 = 1 - 0.6^2)$ for both groups. In loading non-invariant conditions with the amount of non-invariance = 0.2, the residuals of items 8–10 in group B followed $N(\mu = 0, \sigma^2 = 1 - 0.4^2)$. After the data for the latent variable underlying each indicator are generated, I categorized them into ordinal data based on selected thresholds. This process of generating ordinal complete data with a continuous auxiliary variable was similar to Wu, et al. (2015).

3.1.1.1 Population loadings and thresholds in invariant conditions

For invariant conditions, the loadings in both groups are always set to 0.6; thresholds of items were set to be either symmetric (τ_n = -1.3, -0.47, 0.47, 1.3) or asymmetric (τ_n = -0.253, 0.385, 0.842, 1.282). The percentages of the observations per category in the population model were 10%, 22%, 36%, 22%, and 10%, and 40%, 25%, 15%, 10%, and 10%, respectively, for the symmetric and asymmetric conditions. These settings were identical to those of Sass et al. (2014).

3.1.1.2 Population loadings and thresholds in non-invariant conditions

For non-invariant conditions, non-invariance was limited to the last three items in group B (i.e., items 8, 9 and 10 in group B). For a loading non-invariant condition, the loadings of these three items were 0.6 subtracted by 0.2, 0.3, 0.4 or 0.5. In a threshold non-invariant condition, all of the thresholds in the last three items in group B were subtracted by 0.2, 0.3, 0.4 or 0.5. The details of these non-invariant parameters were presented in Table 2.

3.1.1.3 Sample size:

The total sample size was 300 (150 per group), 500 (250 per group) or 1000 (500 per group) to represent small, medium, or large sample sizes. These settings were identical to those of Sass et al. (2014).

3.1.2 Missing data generation and within replication-conditions

After generating 500 complete data sets for each combination of between-replication conditions, I created two incomplete data sets based on each complete dataset with missing data rates of 30% and 50% (in group B), respectively. After that, all datasets are sent to Mplus for ME/I test with the three different strategies (i.e., $FIML_{mvn}$, robust FIML and WLSMV_PD).

3.1.2.1 Missing data conditions

The missing data rates in study 1 were set to be 0%, 30%, and 50%. As mentioned above, only the last three items in group B have missing data. The proportion of the items that contain missing data (3/10, approximately 1/3 of the items in group B) and the missing data rate per item (30% or 50%) are determined based on previous studies (Enders, 2001; Wu et al., 2015).

Missing data were generated as follows. First, the scores of the auxiliary variable were rank-ordered from smallest to largest. The probability of a score being missing wa s then calculated based on the rank order. For example, the probability of having missing data on the 8th item in group B for individual *i* is computed as 1 – (the rank order of the corresponding score of the auxiliary variable/number of the observations in group B). This probability was then compared to a random number k drawn from a uniform distribution, k ~ UNIF(0,1). If the calculated probability is bigger than k, then the 8th item has a missing observation for individual i. This missing data mechanism is MAR. Specifically, as the auxiliary variable score decreases, the probability of the score being missing increases. This process is continued until the desired percentage (30% or 50%) of missing data is achieved for each of the three items, respectively.

In sum, the design factors in study 1 included sample size (300, 600, 1000), locations of non-invariance (invariance in all parameters, non-invariance in loadings, or non-invariance in thresholds), amount of non-invariance (0.2, 0.3, 0.4, 0.5), distributions of the thresholds (symmetric and asymmetric), missing data proportions (0, 30%, 50%), estimators and missing data methods ($FIML_{mvn}$, robust FIML, WLSMV_PD). Except for the last two factors, all others were between-replication factors. In total, there were fifty-four between-replication conditions: 48 non-invariant conditions = sample sizes (3) × locations of non-invariance (2) × amounts of non-invariance (4) × distributions of the thresholds (2), and 6 invariant conditions = sample sizes (3) × distributions of the thresholds (2). For each between-replication condition, 500 datasets were generated using R, and then different amount missing data rates are imposed to the last three indicators in group B. After that, the simulated data sets (complete, 30 % missing datasets, and 50% missing data sets) were sent to Mplus for ME/I tests. Lastly, R is used to analyze the Mplus outputs with the outcome variables described in section 3.1.3.

3.1.2.2 Implementation of the Three Strategies

In all conditions, I conducted ME/I tests with $FIML_{mvn}$, robust FIML, and WLSMV_PD

using Mplus. The example Mplus syntaxes of these methods are presented in Appendix A.

Note that when $FIML_{mvn}$ and robust FIML and WLSMV_PD are used, the auxiliary variable was included in both groups when missing data present, using the saturated correlation model proposed by Graham (2003).

3.1.3 Outcome evaluations

In study 1, I focused on the type I error rate and the power of $\Delta\chi^2$ tests. Type I error rate was calculated for each of the invariance conditions and power was calculated for each of the non-invariance conditions. Both were calculated as the proportion of replications that yielded significant chi-squared different tests ($p < .05$).

Following Rhemtulla et al. (2012), the bias of loading estimates and their corresponding standard errors obtained from configural invariance models were also evaluated with mean relative bias (MRB). The MRB of loadings is defined as $\frac{(\bar{\theta}_{est}-\theta)}{\theta}$, where θ is the population value of loadings and $\bar{\theta}_{est}$ is the average of loading estimates across all replications in a given cell of the condition matrix. The MRB of the standard error of loadings is defined as $\frac{\overline{SE}_{est}-SE_{emp}}{SE_{emp}}$, where $\overline{SE}_{est}$ was the average estimate standard error in a given cell. SE_{emp} is the empirical standard deviation of the associated parameter, which can be considered as a proxy of true standard errors.

3.2 Study 2

The simulation conditions for study 2 were determined based on the studies conducted by Jung and Yoon (2016), Sass et al. (2014), and Wu et al. (2015). Specifically, I use a two-group, single factor CFA model that has six five-point indicators per group as the basis for the population models in study 2. This population model is presented in Figure 2. Note that instead of using the 10-indicator model in study 1, I used a six-indicator model in study 2. The main reason is that the six-indicator model is comparable to those used in Jung & Yoon (2016) and other studies on model modifications (e.g., Yoon & Kim, 2014), so that I can compare my results to those from previous research. Group A was used as the reference group, where data were always complete and all the model parameters were fixed across all conditions; group B was considered as the focus group, where the population loadings or the second thresholds in items 2

and 4 were changed in non-invariant conditions. Specifically, all items in group A and items 1, 3, 5, and 6 in group B have factor loadings fixed at 0.7 and thresholds fixed at -1.3, -0.47, 0.47, and 1.3 across all conditions. The loadings and the “second” thresholds (originally set at -0.47 in invariant conditions) in items 2 and 4 in group B were subtracted by certain values to create the non-invariance. The amounts of non-invariance are varied depending on the patterns of non-invariance. Missing data are imposed only on items 2 and 4 .

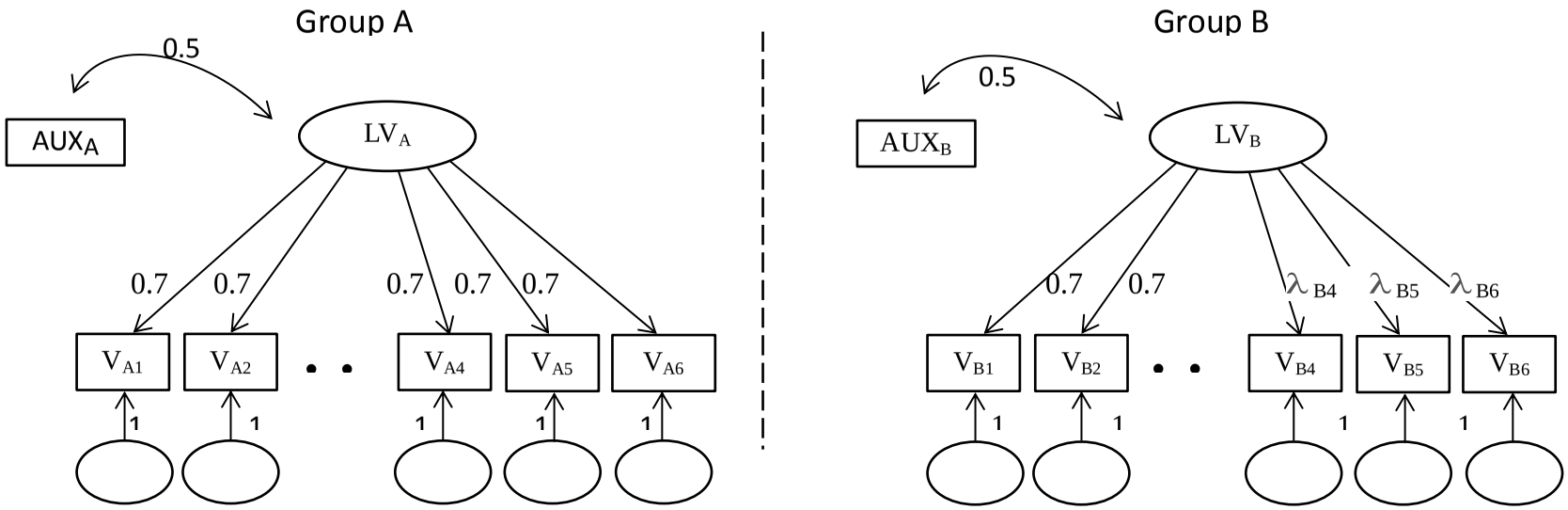

Figure 2. The population model for study 2

Note: AUX_A and AUX_B represent the auxiliary variables for groups A and B. LV_A and LV_B represent the latent variables for groups A and B. LV_A(LV_B) and AUX_A(AUX_B) jointly follow a $MVN(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix})$.

The designed factors in study 2 are summarized as follows:

1. Sample size (N = 250, 500, or 1000 per group),
2. Locations of invariance (invariant, non-invariance loadings, non-invariance thresholds),
3. Patterns of non-invariance (small, large, mixed-size, or non-uniform invariance). The specific parameter values of these patterns are discussed in section 3.2.1.2 and presented in Table 3.
4. Missing data proportions (0, 30%, 50%)
5. Four specification search methods (backward MFI methods with FIML, robust FIML, or WLSMV_PD, and the forward CI search method with WLSMV_MI).

Factors 1 to 3 were between-replication factors; factors 4 and 5 were within-replication factors. For each combination of the between-replication conditions, I generated 500 datasets. The general process involved the same five steps as those in study 1 except that in step 4, the datasets were sent to Mplus for specification search analyses rather than ME/I tests. The simulated conditions are described in detail below.

Table 3. *Model parameters of different patterns of non-invariance in study 2*

Parameter	Group A	Group B				
		Baseline (item 1,3,5,6 in group B)	Small difference	Large difference	Mixed-size difference	Non-uniform difference
Loading in item 1	.7	.7	.7	.7	.7	.7
Loading in item 2	.7	.7	**.5**	**.3**	**.4**	**.4**
Loading in item 3	.7	.7	.7	.7	.7	.7
Loading in item 4	.7	.7	**.5**	**.3**	**.2**	**1.0**
Loading in item 5	.7	.7	.7	.7	.7	.7
Loading in item 6	.7	.7	.7	.7	.7	.7
Thresholds in item 1	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)
Thresholds in item 2	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, **-0.67,** 0.47, 1.3)	(-1.3, **-0.87**, 0.47, 1.3)	(-1.3, **-0.77**, 0.47, 1.3)	(-1.3, **-0.77**, 0.47, 1.3)
Thresholds in item 3	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)
Thresholds in item 4	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, **-0.67,** 0.47, 1.3)	(-1.3, **-0.87**, 0.47, 1.3)	(-1.3, **-0.97**, 0.47, 1.3)	(-1.3, **-0.17**, 0.47, 1.3)
Thresholds in item 5	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)
Thresholds in item 6	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)	(-1.3, -0.47, 0.47, 1.3)

Note: Locations of non-invariance are set to either loadings or thresholds in items 2 and 4 in group B. Non-invariance occurs on either loadings or thresholds.

3.2.1 Complete data generation and between replication conditions for study 2

Same as study 1, I assume that there is a continuous and normal distributed latent variable underlying each indicator. To generate the ordinal data, bivariate normal data are generated with the MASS package in R for each group (R core team, 2016). In each group, I again use the "mvnorm" function to generate a standardized bivariate normal distributed data X, where the covariance is set to 0.5 (i.e., $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim MVN(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$) (i.e., the means and variances of factors are always fixed at 0 and 1 respectively), where X_1 is used as factor scores and X_2 is used as a continuous auxiliary variable that can later be used to generate missing data. The factor scores were multiplied to the population loadings to create the continuous latent response variate underlying each ordinal indicator. After that, for each latent response variate, a normally distributed error term ($\varepsilon \sim N(\mu = 0, \sigma^2 = 1 - loading^2)$ was added to it. Specifically, for items that do not have non-invariant loadings, the residual variances are all fixed at $1 - 0.7^2 = 0.51$. For loading non-invariant conditions, the residual variances of items 2 and 4 in group B are subtracted by a certain value, depending on the patterns of non-invariance.

3.2.1.1 Population loadings and thresholds in invariant conditions

In study 2, for invariant conditions, population loadings in both groups were set to 0.7, following Jung & Yoon (2016); thresholds are set to (τ_n = -1.3, -0.47, 0.47, 1.3), borrowing the symmetric thresholds used in Sass et al. (2014).

3.2.1.2 Population loadings and thresholds in non-invariant conditions

For the non-invariant conditions, the non-invariances are limited to item 2 and item 4 in group B, either on their loadings or thresholds. In study 2, I followed Jung & Yoon (2016) to create four different patterns of non-invariance. These patterns are named small difference, large difference, mixed-size difference, and non-uniform difference. The parameter values for each of these patterns are presented in Table 2. For loading non-invariant conditions, the population loadings of items 2 and 4 in group B (originally set to 0.7, 0.7) are subtracted by (0.2, 0.2), (0.4, 0.4) (0,3. 0.5) and (0.3, -0.3) in the small difference, large difference, mixed-size difference and non-uniform difference conditions. For example, in a mixed-size difference loading non-invariant condition, the factor loadings of item 2 and 4 in group B are set at 0.7 – 0.3= 0.4 and 0.7 – 0.5 = 0.2, respectively. Similarly, the threshold non-invariant conditions were

generated by imposing the differ types of non-invariance on the second thresholds of items 2 and 4 in group B. The second thresholds in these two items in group B (originally set to be -0.47. -0.47) were subtracted by (0.2, 0.2), (0.4, 0.4) (0,3. 0.5) and (0.3, -0.3) in the small difference, large difference, mixed-size difference and non-uniform difference conditions, respectively. In a non-uniform, threshold non-invariant condition, the thresholds of item 2 and item 4 are set at τ_n = -1.3, -0.47-0.3, 0.47, 1.3, and -1.3, -0.47+0.3, 0.47, 1.3, respectively. The specific population parameters in these non-invariant conditions are presented in Table 2.

3.2.1.3 Sample size

The total sample size in study 2 was set to 500 (250 per group), 1000 (500 per group), or 2000 (1000 per group). These settings were identical to those in Jung & Yoon (2016).

3.2.2 Missing data generation and within-replication conditions for study 2

The within-replication factors included missing data proportions and the specification search methods.

3.2.2.1 Missing data proportions

The missing data rate was set to be 0%, 30%, or 50%. Note that missing data are only imposed on items 2 and 4 in group B. Missingness was determined by the auxiliary variable and was generated in the same way as in study 1.

3.2.2.2 Implementation of Different Strategies

For each data set in study 2, I conducted the backward specification search with modification indices obtained from $FIML_{mvn}$, robust FIML, and WLSMV_PD. Similar to study 1, the auxiliary variable was included in both groups by using the saturated correlation model proposed by Graham (2003) when $FIML_{mvn,}$ robust FIML and WLSMV_PD were used. The backward specification search procedure starts with the metric or scalar invariance model, depending on the location of the non-invariance. The modification indices are obtained by using the Mplus command “OUTPUT: MODINDICES (cutoff value)”. The cutoff values were set to be 3.841 and 6.635, to represent the criteria of 0.05 and 0.01 significance levels of the chi-squared statistic with 1 degree of freedom. For example, a backward specification search for a loading invariance with the 0.05 significance level will start by fitting a metric invariance model to the data. Among all the modification indices on equality constraints in this metric model, the largest one above 3.841 will be released. A partial metric invariance model with one pair of free

loadings will then be re-estimated to generate the second set of modification indices, which can tell us the next equality constraint to be released. The procedure continues until all modification indices on equality constraint remaining in the partial invariance model are smaller than 3.841. An example of the Mplus syntax of backward MFI search for non-invariant thresholds is presented in Appendix A.

Besides the backward search methods, I also conducted the forward CI search method with WLSMV_MI through Mplus. Specifically, the configural/metric invariance models was fit to the imputed data sets with the MODEL CONSTRAINT commands. Example syntax for searching non-invariant loadings and non-invariant thresholds with confidence intervals is presented in Appendix A. Note that in the forward specification search, I always used WLSMV as the estimator. Furthermore, the number of imputations is set to be 20, following Enders (2010).

In sum, in study 2, the designed factors included sample sizes (500, 1000, 2000), locations of non-invariant items (all invariant, non-invariant in loadings or thresholds), patterns of non-invariance (small, large, mix, or non-uniform), missing data proportions (0, 30%, 50%), and specification methods (backward MFI search obtained from $FIML_{mvn}$, robust FIML, WLSMV_PD, and the forward CI research with WLSMV_MI). Except for missing data proportions and search methods, all of the factors are between-replication conditions. There were 30 between-replication conditions (Note: 24 non-invariant conditions = sample sizes (3) × occurrence of non-invariance (2) × patterns of the non-invariance (4) × distribution of the thresholds (2); 6 invariant conditions: sample size (3) × loading or threshold invariance (2)). For each cell of these between-replication conditions, 500 datasets are generated using R, and then different amounts of missing data are imposed to create incomplete data. After that, the data sets (complete, 30 % missing data set, and 50% missing data set) were sent to Mplus for specification searches (with four different methods). Lastly, R was used to analyze Mplus outputs with the outcome variables mentioned as follows.

3.2.3 Outcome evaluations

In study 2, I focused on the capability of different specification search methods to correctly identify non-invariant items. I examine three kinds of outcome variables, following Jung & Yoon (2016). They are (1) perfect recovery rates, (2) model level type I and type II error rates, and (3) item level power and type I error rate. These outcome variables are defined as follows.

3.2.3.1 Perfect recovery rates

A perfect recovery rate can be considered as a criterion that is similar to but more rigorous than power. A perfect recovery means that all non-invariant and invariant items are correctly identified in a replication in the final partial invariance model. In study 2, I calculated the perfect recovery rate of each method for each non-invariant condition. For example, in a non-invariant condition, if there are K replications where a method successfully identifies all the non-invariant and invariant items, then the perfect recovery rate of this method in this condition was calculated as k/500.

3.2.3.2 Model-level type I and type II error rates

Following Jung & Yoon (2016), in study 2, I defined the model level type I error as the probability of misidentifying any invariant items as non-invariant in the final partial invariance model. This probability is estimated for both invariant and non-invariant conditions across the 500 replications within each condition. For example, if there were k replications with any misspecified non-invariant items within that condition, the model level type I error rate was calculated as k/500. Similarly, model level type II error is defined as the probability of misidentifying any non-invariant item as an invariant item. For example, in a non-invariant condition with a specific missing data rate (e.g., 30%), the model level type II error rate for a method will be calculated as k/500, where k is the number of replications where the final partial model has misidentified any non-invariant items as invariant (within 500 replications).

3.2.3.3 Item-level type I error rates and power

Besides model level outcomes, I also compared these methods at the item level. Specifically, I calculated the item level power as the proportion of correctly identified non-invariant items (pairs) over the number of non-invariant items (pairs) multiplied by 500 replications. Similarly, the item-level type I error rate was calculated as the proportion of identified invariant items (pairs) over the number of the tested non-invariant items (pairs) multiplied by 500 replications.

Using a non-invariant condition as an example, there were two pairs of non-invariant items (items 2 and 4) and three pairs of invariant items (items 3, 5, and 6). Consequently, the denominator used to compute the item level power is 2 (# non-invariant items) × 500 (# of replications) = 1000, and the numerator was the number of paired items that are correctly identified as non-invariant across 500 replications. On the other hand, the denominator of the

item-level type I error rate in a non-invariant condition was 3 (invariant items) × 500 (# of replications) = 1500, and the numerator is the number of the paired items that had been misidentified as non-invariant across 500 replications. Note that because non-invariant conditions contain invariant and non-invariant items simultaneously, I calculated both the item-level type I error rate and power for non-invariant conditions.

Chapter 4─Simulation 1 Results

When discussing the results of simulation 1 and 2, I used the terms $FIML_{mvn}$, robust FIML, WLSMV_PD and WLSMV_MI to indicate different methods for ME/I testing and specification search. For conditions with complete data, auxiliary variables were not included into analyses. The terms $FIML_{mvn}$, robust FIML and WLSMV_PD simply indicated that the results obtained from the ML_{mvn}, MLR and WLSMV estimators respectively. WLSMV_MI also indicated that the results of specification search based on confidence intervals by using WLSMV with complete data.

4.1 Non-Convergence and Improper Solutions in Study 1

All replications with the FIML methods converged. Only 9 out of 81,000 replications had convergence problems with WLSMV_PD. As for the rates of improper solution for the loading estimates, FIML and robust FIML were more likely than WLSMV_PD to produce improper estimates (i.e., standardized loading bigger than 1), especially with a small sample size and high missing data rate. The results were summarized in Appendix B (see Figure B1). Improper solution rates were low for all methods (less than 10%), and they decreased as sample size increased. With N > 300, they were less than 2%. The replications with improper solutions were excluded from the rest of the analyses.

4.2 Type I Error Rates of $\Delta\chi^2$ Tests

The results for type I error rates were summarized in Table 4. In general, the two FIML methods outperformed WLSMV_PD in controlling type I error rates at the nominal level. Similar result patterns were observed for $FIML_{mvn}$ and robust FIML, except that $FIML_{mvn}$ overly controlled type I error rates where the sample size was 600. The type I error rates obtained from robust FIML all fell within the acceptable range (0.03 – 0.069), except for the conditions that the sample size was small, the missing data rate was 50%, and the thresholds were asymmetric (α =0.018).

Table 4. *Type I error rate of* $\Delta\chi^2$ *tests*

Method	Thresholds	N=300			N=600			N=1000		
		complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	Asymmetric	0.043	0.032	**0.021**	0.054	0.042	0.049	0.046	0.042	0.060
rFIML		0.034	0.032	**0.018**	0.058	0.042	0.055	0.058	0.048	0.058
WLSMVPD		0.062	**0.072**	**0.130**	0.058	**0.094**	**0.250**	0.046	**0.108**	**0.388**
FIML	Symmetric	0.040	0.036	0.046	**0.028**	**0.026**	0.032	0.032	0.032	0.054
rFIML		0.056	0.050	0.054	0.040	0.040	0.040	0.056	0.044	0.064
WLSMVPD		0.062	**0.088**	**0.180**	0.036	**0.092**	**0.292**	0.052	**0.122**	**0.550**

The values fall out of the acceptable range (0.03 – 0.069) were highlighted. rFIML indicates robust FIML.

In comparison, the type I error rates from WLSMV_PD were highly influenced by missing data rate. With complete data, the type I error rates from WLSMV_PD were all acceptable (.036 - .069). However, with missing data, the type I error rates from WLSMV_PD were inflated (> .069), especially when the sample size and missing data rate were both large. For example, with 50% missing data and N = 1000, the type I error rate could be as large as 0.55.

4.3 Power of $\Delta\chi^2$ Tests

The results for power to detect non-invariance in loadings and thresholds are summarized in Figures 3 and 4. Give that the results were very similar between the conditions with symmetric and asymmetric thresholds, I presented only the results for the conditions with symmetric thresholds. The results for the asymmetric conditions can be found in Appendix B (see Figures B2 and B3).

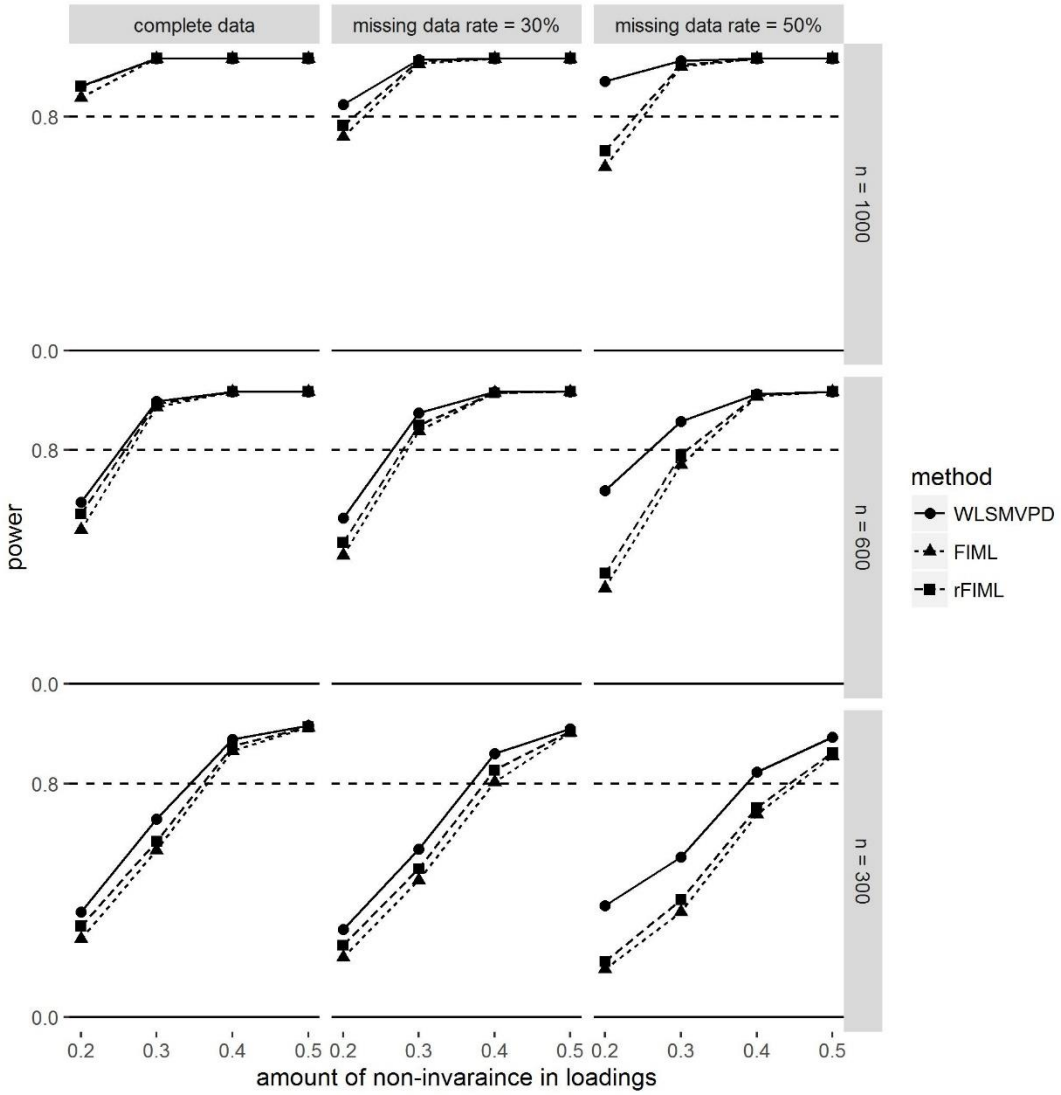

Figure 3. Power* of the $\Delta\chi^2$ tests on detecting non-invariant loadings when thresholds are symmetric.

Note: FIML: continuous full information likelihood method, rFIML: robust continuous full information likelihood method, WLSMVPD: weighted least squares means and variance adjusted estimators plus pairwise deletion.

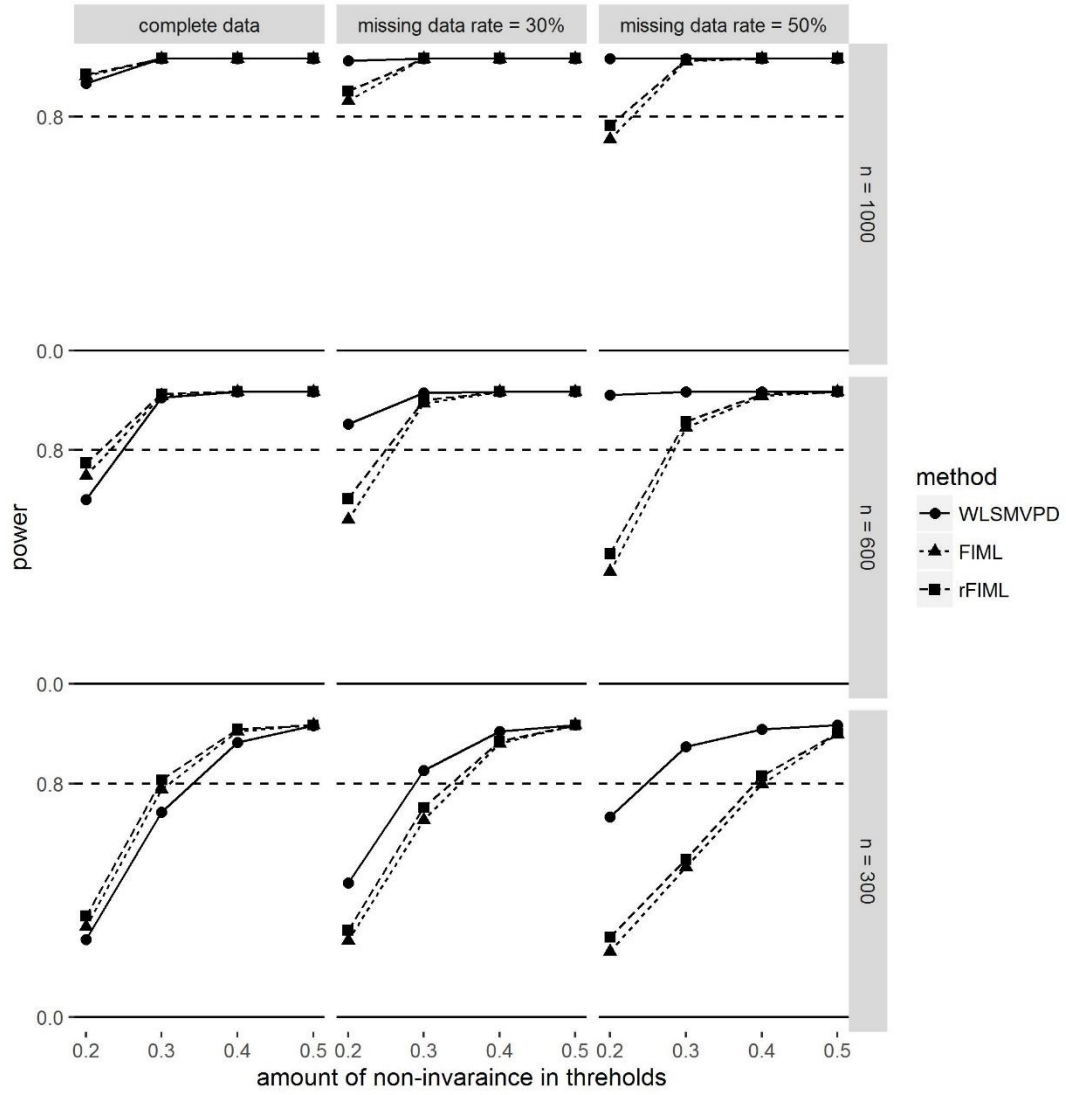

Figure 4. Power* of the $\Delta\chi^2$ tests on detecting non-invariant thresholds when thresholds are symmetric

Note: FIML: continuous full information likelihood method, rFIML: robust continuous full information likelihood method, WLSMVPD: weighted least squares means and variance adjusted estimators plus pairwise deletion.

It is well known that power will increase as sample size or effect size increases. When the sample size was large (n = 1000), all methods had sufficient power (> 0.8) to detect non-invariance, except for very few conditions where the amount of non-invariance was small (amount of non-invariance = 0.2). Similarly, when the amount of non-invariance was sufficiently

large (≥ .40), all methods had sufficient power to detect non-invariance regardless of sample size. In addition, holding the other factors constant, increase in the missing data rate resulted in a decrease in the power of $\Delta\chi^2$ for all methods.

Both $FIML_{mvn}$ and robust FIML had comparable power rates across all conditions. They both differed from WLSMV_PD. How they were different depended on the locations of non-invariance and the missing data rate. As shown in Figure 3, when non-invariance occurred in the loadings, WLSMV_PD had more power than the FIML methods to detect non-invariance, regardless of whether there were missing data or not. When non-invariance occurred in the thresholds the power rates from the two FIML methods were close to (or even slightly outperformed) that of WLSMV_PD with complete data, especially when the amount of non-invariance was small. With the presence of missing data, however, WLSMV_PD showed higher power rates. However, the high power of WLSMV_PD present in these figures could just be the side effect (spurious power) of the severely inflated type I error rate presented in Table 4.

4.4 Relative Biases of Loading Estimates

For ease of presentation, I separated the items into two groups. The first group contains complete items for which the data were always complete. These items included all items in group A and items 1 – 7 in group B. The second group contains three items that had missing data in some of the conditions (i.e., items 8 -10 in group B). I referred to these items as incomplete items. Given that the results for the complete items were consistent across groups A and B, I only presented the results for the complete items in group B. The results related to items in group A can be found in Appendix B.

In addition, since the relative biases were consistent for the complete items or for the incomplete items, I reported the mean relative biases (MRB) for these items in Figures 5 and 6. Note that because the point estimates were identical between FIML and robust FIML, only the results from robust FIML were reported. In addition, because the location of non-invariance and amount of non-invariance did not affect the relative performance of the methods, I collapsed the results across the two factors in Figures 5 and 6.

As can be seen in Figure 5, the MRBs of loadings from the FIML methods for complete items were mainly determined by the thresholds distributions. When the thresholds were symmetric, the MRBs from the three methods were all within the acceptable range (i.e., |RB| <

0.1), regardless of the missing data proportions. However, when the thresholds were asymmetric, the MRBs obtained from the FIML methods became substantial. In contrast, the MRBs from WLSMV_PD were not increased by the asymmetric thresholds. Similar patterns were observed for incomplete items (see Figure 6), and missing data rates had a minimal impact on the MRBs for the three methods.

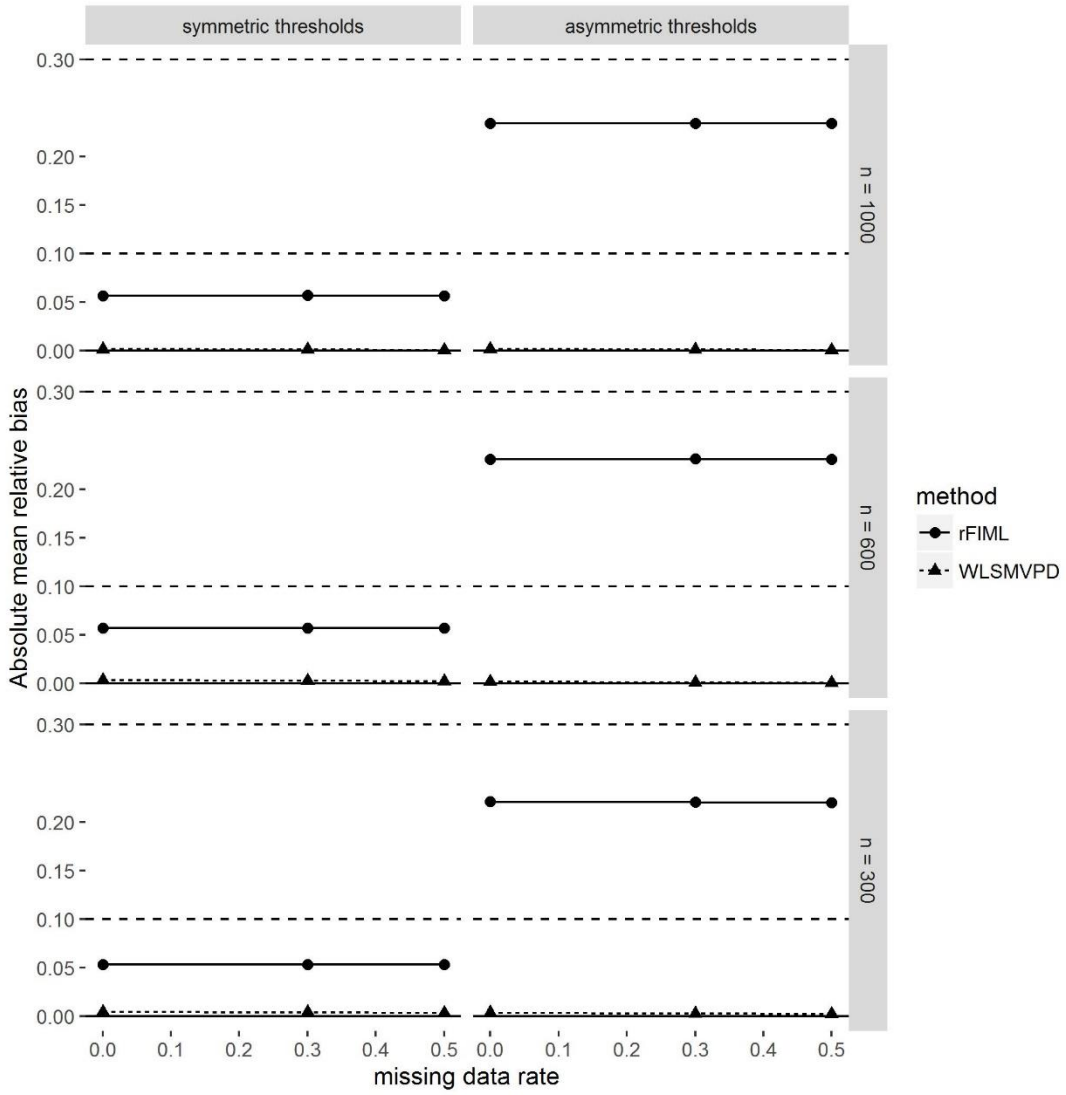

Figure 5. Absolute mean relative bias estimates across complete items in group B.

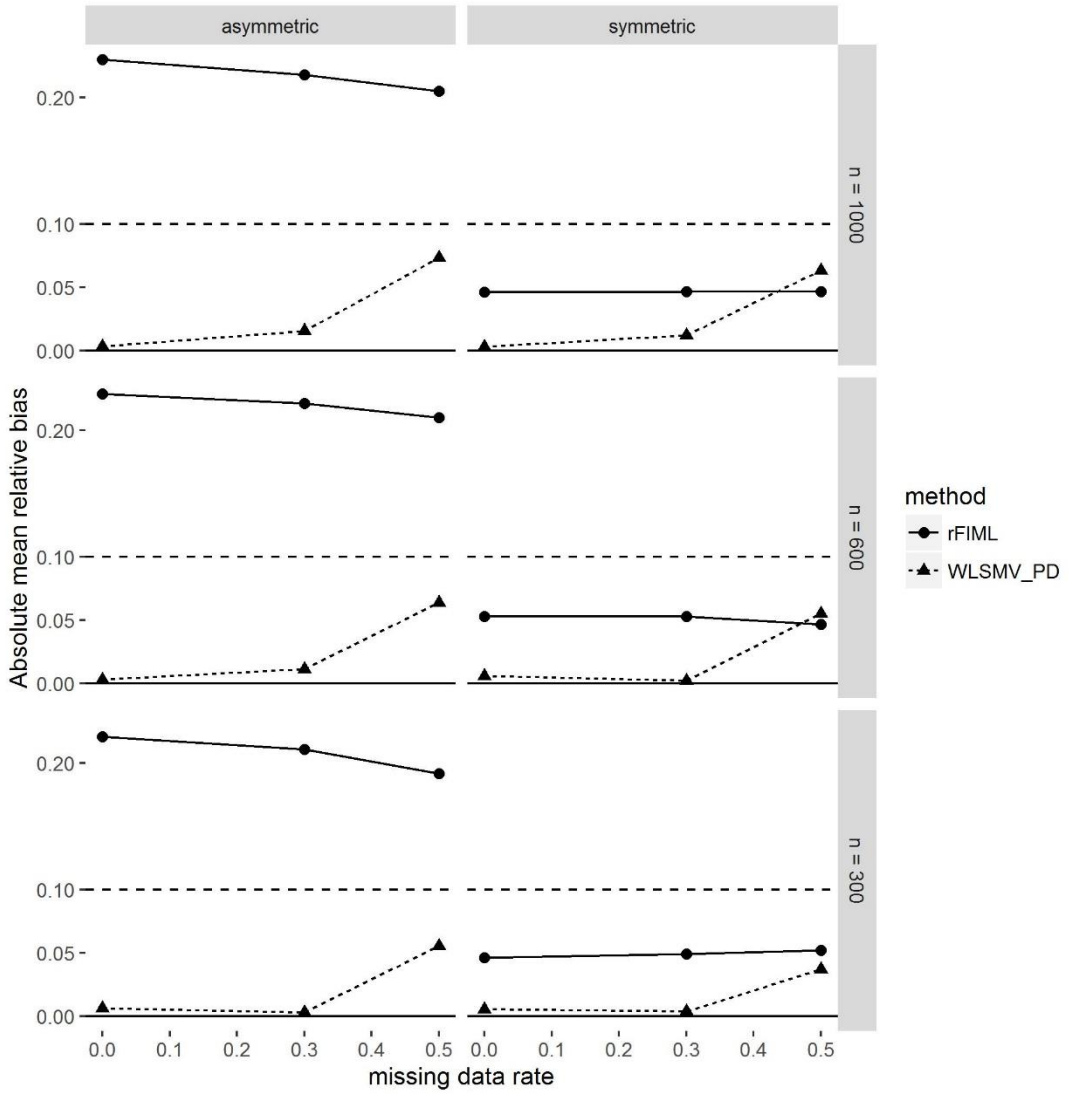

Figure 6. Absolute mean relative bias estimates across incomplete items in group B.

4.5 Relative Biases in Estimates of Standard Error

As for the standard errors (SE) for loading estimates, the MRBs of SE obtained from all methods were all within the acceptable range (|RB| < 0.1, see Appendix B Tables B7 to B10) for complete items. For incomplete items, however, substantial biases in the SEs were observed under certain conditions for all three methods. The results for symmetric and asymmetric thresholds were summarized in Tables 5 and 6, respectively. As can be seen in Table 5, when the thresholds were symmetric, the MRBs of standard errors from the FIML methods were generally acceptable, while the MRBs from WLSMV_PD were substantially biased when missing data presented with a small sample size (n = 300).

Table 5. *Mean relative biases of standard errors for loadings across incomplete items in group B with symmetric thresholds*

Estimator	DIF_T	DIF_L	N = 300			N = 600			N = 1000		
			complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.082	0.043	0.044	0.054	0.051	0.058	0.052	0.048	0.011
	0.2	0	0.059	0.068	0.023	0.023	0.034	0.015	0.083	0.087	0.050
	0.3	0	0.057	0.043	0.048	0.046	0.078	0.033	0.069	0.079	0.037
	0.4	0	0.069	0.044	0.024	0.033	0.008	0.007	0.064	0.064	0.037
	0.5	0	0.080	0.088	0.056	0.042	0.044	0.051	0.049	0.053	0.057
	0	0.2	0.026	0.004	0.029	0.041	0.026	0.013	0.045	0.052	0.008
	0	0.3	0.016	0.033	-0.011	-0.005	-0.005	0.005	0.024	0.010	0.023
	0	0.4	-0.001	-0.011	-0.035	0.014	0.002	0.023	0.015	-0.005	0.005
	0	0.5	-0.031	-0.037	-0.018	-0.003	-0.013	-0.027	-0.011	-0.006	-0.016
rFIML	0	0	0.030	-0.003	0.006	-0.004	-0.003	0.010	-0.007	-0.006	-0.036
	0.2	0	0.007	0.019	-0.018	-0.033	-0.020	-0.030	0.021	0.029	-0.004
	0.3	0	0.003	-0.005	0.003	-0.013	0.023	-0.015	0.008	0.021	-0.014
	0.4	0	0.016	-0.004	-0.013	-0.024	-0.042	-0.039	0.005	0.006	-0.012
	0.5	0	0.027	0.040	0.017	-0.015	-0.011	0.002	-0.010	-0.002	0.005
	0	0.2	0.017	-0.004	0.026	0.020	0.007	-0.002	0.021	0.028	-0.013
	0	0.3	0.018	0.034	-0.009	-0.011	-0.010	0.000	0.014	0.001	0.015
	0	0.4	0.011	-0.001	-0.024	0.017	0.005	0.027	0.014	-0.006	0.004
	0	0.5	-0.014	-0.019	0.001	0.003	-0.009	-0.022	-0.007	-0.003	-0.014
WLSMVPD	0	0	-0.043	-0.089	**-0.134**	-0.037	-0.075	-0.058	-0.022	-0.033	-0.064
	0.2	0	-0.049	-0.063	**-0.136**	-0.055	-0.046	-0.065	-0.013	-0.004	-0.021
	0.3	0	-0.066	-0.094	**-0.124**	-0.048	-0.026	-0.086	-0.019	-0.021	-0.039
	0.4	0	-0.029	-0.089	**-0.125**	-0.055	-0.081	-0.090	-0.022	-0.035	-0.049
	0.5	0	-0.052	-0.078	**-0.121**	-0.055	-0.064	-0.058	-0.028	-0.037	-0.046
	0	0.2	-0.072	**-0.108**	**-0.110**	-0.020	-0.041	-0.056	-0.008	0.005	-0.043
	0	0.3	-0.062	-0.070	**-0.132**	-0.047	-0.055	-0.059	-0.020	-0.034	-0.038
	0	0.4	-0.074	**-0.105**	**-0.138**	-0.022	-0.041	-0.034	-0.009	-0.029	-0.024
	0	0.5	-0.094	**-0.113**	**-0.129**	-0.039	-0.052	-0.082	-0.029	-0.026	-0.047

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, miss: missing

Table 6. *Mean relative biases of standard errors for loadings across incomplete items in group B with asymmetric thresholds*

Estimator	DIF_T	DIF_L	N = 300			N = 600			N = 1000		
			complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.023	0.059	0.076	0.043	0.052	0.090	0.009	0.018	0.020
	0.2	0	**0.115**	**0.185**	**0.185**	**0.105**	**0.121**	**0.173**	0.039	0.040	0.060
	0.3	0	**0.169**	**0.188**	**0.267**	**0.106**	**0.107**	**0.148**	**0.102**	**0.106**	**0.107**
	0.4	0	**0.189**	**0.182**	**0.229**	**0.171**	**0.184**	**0.162**	**0.164**	**0.172**	**0.187**
	0.5	0	**0.222**	**0.239**	**0.223**	**0.141**	**0.156**	**0.134**	**0.117**	**0.122**	**0.105**
	0	0.2	-0.030	-0.009	-0.037	0.036	0.037	0.047	0.014	0.014	0.007
	0	0.3	-0.045	-0.043	-0.021	-0.038	-0.030	-0.042	0.012	0.012	0.002
	0	0.4	-0.017	0.004	-0.035	-0.041	-0.046	-0.026	-0.032	-0.025	-0.020
	0	0.5	-0.041	-0.027	-0.052	-0.024	-0.017	-0.034	0.015	0.003	-0.004
rFIML	0	0	0.025	0.049	0.052	0.030	0.020	0.041	-0.009	-0.018	-0.034
	0.2	0	0.056	**0.118**	**0.118**	0.036	0.042	0.085	-0.032	-0.041	-0.030
	0.3	0	0.086	**0.105**	**0.177**	0.013	0.009	0.045	0.005	0.002	0.000
	0.4	0	0.084	0.081	**0.134**	0.054	0.063	0.048	0.043	0.047	0.062
	0.5	0	0.099	**0.121**	**0.122**	0.010	0.026	0.018	-0.015	-0.010	-0.018
	0	0.2	0.007	0.022	-0.014	0.060	0.051	0.050	0.032	0.023	0.007
	0	0.3	-0.001	-0.006	0.012	-0.010	-0.007	-0.026	0.035	0.029	0.011
	0	0.4	0.028	0.045	0.001	-0.012	-0.022	-0.005	-0.008	-0.005	-0.005
	0	0.5	-0.006	0.006	-0.023	0.001	0.006	-0.012	0.033	0.018	0.008
WLSMVPD	0	0	-0.052	-0.067	**-0.145**	-0.002	-0.015	-0.035	-0.030	-0.047	-0.067
	0.2	0	-0.038	-0.038	**-0.111**	-0.005	-0.016	-0.034	-0.047	-0.066	-0.062
	0.3	0	-0.044	-0.083	-0.087	-0.038	-0.051	-0.057	-0.017	-0.021	-0.034
	0.4	0	-0.048	-0.075	-0.072	-0.001	0.011	-0.047	0.018	0.001	-0.014
	0.5	0	-0.043	-0.066	-0.083	-0.021	-0.037	-0.082	-0.035	-0.037	-0.043
	0	0.2	-0.074	-0.082	**-0.142**	0.011	-0.019	-0.025	0.009	0.006	-0.026
	0	0.3	-0.077	-0.086	**-0.116**	-0.052	-0.052	-0.078	0.006	-0.005	-0.018
	0	0.4	-0.059	-0.060	**-0.107**	-0.060	-0.071	-0.058	-0.030	-0.029	-0.040
	0	0.5	-0.080	-0.085	**-0.139**	-0.040	-0.046	-0.076	0.010	-0.005	-0.018

Note: rFIML= robust FIML, WLSMV_PD = mean and variance adjusted weight least squared with pairwise deletion, DIF_T = amount of non-invariance in thresholds, DIF_L = amount of non-invariance in loadings, miss: missing.

When thresholds were asymmetric (see Table 6), the MRBs of SE obtained from FIML became substantial in all conditions where non-invariance occurred in thresholds. The standard errors from robust FIML were much more robust to asymmetric thresholds. They were substantively biased only in a few conditions where the sample size was small and missing data occurred along with non-invariant thresholds. The results from WLSMV_PD were very similar to that in Table 5. The standard errors were substantially biased when the sample size was small and the missing data rate was 50%.

4.6 Influence of unbalance sample sizes

I only considered balanced group sizes in study 1 originally. Given that group sizes are very likely to be unbalanced in practice and this unbalance may have substantial influence on the analysis result especially given that the total sample size is small, I added a small simulation study to examine the effect of unbalanced group sizes under small sample sizes. Specifically, for n = 300, I added a condition where the reference group proportion is 66.7% and the focal group proportion is 33.3%. I considered only symmetric thresholds for the ordinal data in this case.

The results from the simulation are presented in Tables 7, 8, and 9. As shown in Table 7, unbalance group sizes inflated the type I error rate (comparing the upper panel to the lower panel in Table 7). In addition, unbalance group sizes slightly decreased the power of $\Delta\chi^2$ test (see Tables 8 and 9).

Table 7. *Type I error rates of* $\Delta\chi^2$ *tests in conditions with balanced and unbalanced group sizes*

	Method	Complete data	30 % Missing rate	50 % Missing rate
Balanced	FIML	0.040	0.036	0.046
	rFIML	0.056	0.050	0.054
	WLSMVPD	0.062	0.088	0.180
unbalanced	FIML	0.052	0.038	0.037
	rFIML	0.072	0.068	0.049
	WLSMVPD	0.074	0.090	0.156

rFIML indicates robust FIML

Table 8. *Power of* $\Delta\chi^2$ *tests in loading non-invariant conditions with balanced and unbalanced group sizes*

method	Missing data rates	Amount of non-invariance	Balanced group size	unbalanced group sizes
FIML	complete	0.2	**0.268**	0.236
	30%	0.2	**0.204**	0.174
	50%	0.2	0.164	**0.168**
rFIML	complete	0.2	**0.312**	0.262
	30%	0.2	**0.246**	0.212
	50%	0.2	**0.190**	0.186
WLSMVPD	complete	0.2	**0.360**	0.314
	30%	0.2	**0.300**	0.270
	50%	0.2	**0.382**	0.314
FIML	complete	0.3	**0.572**	0.548
	30%	0.3	**0.468**	0.420
	50%	0.3	**0.360**	0.328
rFIML	complete	0.3	**0.602**	0.568
	30%	0.3	**0.508**	0.446
	50%	0.3	**0.402**	0.362
WLSMVPD	complete	0.3	**0.678**	0.646
	30%	0.3	**0.576**	0.516
	50%	0.3	**0.548**	0.524
FIML	complete	0.4	**0.912**	0.790
	30%	0.4	**0.804**	0.684
	50%	0.4	**0.694**	0.548
rFIML	complete	0.4	**0.928**	0.802
	30%	0.4	**0.846**	0.692
	50%	0.4	**0.718**	0.564
WLSMVPD	complete	0.4	**0.952**	0.854
	30%	0.4	**0.902**	0.780
	50%	0.4	**0.840**	0.710
FIML	complete	0.5	**0.992**	0.946
	30%	0.5	**0.974**	0.866
	50%	0.5	**0.894**	0.756
rFIML	complete	0.5	**0.994**	0.948
	30%	0.5	**0.978**	0.866
	50%	0.5	**0.906**	0.760
WLSMVPD	complete	0.5	**0.998**	0.980
	30%	0.5	**0.988**	0.942
	50%	0.5	**0.958**	0.878

Note. The higher power(s) between the balanced and unbalanced groups size conditions are highlighted in bold. rFIML indicates robust FIML.

Table 9. *Power of* $\Delta\chi^2$ *tests in threshold non-invariant conditions with balanced and unbalanced group sizes*

method	Missing data rates	Amount of non-invariance	Balanced group sizes	unbalanced group sizes
FIML	complete	0.2	0.310	**0.326**
	30%	0.2	**0.262**	0.250
	50%	0.2	**0.224**	0.170
rFIML	complete	0.2	0.346	**0.376**
	30%	0.2	**0.298**	0.284
	50%	0.2	**0.274**	0.202
WLSMVPD	complete	0.2	0.266	**0.279**
	30%	0.2	**0.460**	0.434
	50%	0.2	**0.686**	0.574
FIML	complete	0.3	**0.780**	0.718
	30%	0.3	**0.674**	0.552
	50%	0.3	**0.514**	0.452
rFIML	complete	0.3	**0.812**	0.760
	30%	0.3	**0.718**	0.594
	50%	0.3	**0.540**	0.456
WLSMVPD	complete	0.3	**0.702**	0.663
	30%	0.3	**0.844**	0.750
	50%	0.3	**0.926**	0.842
FIML	complete	0.4	**0.978**	0.958
	30%	0.4	**0.936**	0.868
	50%	0.4	**0.798**	0.646
rFIML	complete	0.4	**0.986**	0.962
	30%	0.4	**0.944**	0.882
	50%	0.4	**0.826**	0.674
WLSMVPD	complete	0.4	**0.940**	0.912
	30%	0.4	**0.978**	0.930
	50%	0.4	**0.986**	0.954
FIML	complete	0.5	**1.000**	**1.000**
	30%	0.5	**1.000**	0.980
	50%	0.5	**0.970**	0.898
rFIML	complete	0.5	**1.000**	**1.000**
	30%	0.5	**1.000**	0.986
	50%	0.5	**0.972**	0.904
WLSMVPD	complete	0.5	**0.998**	0.993
	30%	0.5	**1.000**	0.994
	50%	0.5	**1.000**	0.996

Note. The higher power(s) between the balanced and unbalanced group sizes conditions are highlighted in bold. rFIML indicates robust FIML.

Chapter 5─Simulation 2 Results

In study 2, I did not encounter convergence problems when conducting specification searches with $FIML_{mvn}$, robust FIML, or WLSMV_MI; however, among the 90,000 models fitted with WLSMV_PD, 122 models failed to converge. These non-convergences all occurred for loading non-invariant conditions, especially for the conditions that contained at least one population loading equal to or lower than 0.3 (120/122). The non-convergence rates for these conditions ranged from 0.06% to 2.5%. The results of these models were excluded from the following analyses.

5. 1 Results for loading invariant conditions

The model-level type I error rates obtained from loading invariant conditions were presented in Tables 10 and 11. As mentioned earlier, in these conditions, specification searches were conducted on MFI of the 5 equality constraints on loadings in the metric invariance model for backward methods (i.e., $FIML_{mvn,}$ robust FIML, and WLSMV_PD), and confidence intervals of the configural invariance model for the forward method (i.e., WLSMV_MI). Recalled that the model-level type I error is defined as any invariant pair of loadings being misidentified as non-invariant in the final model.

Table 10. *Basal model-level type I error rates of methods in loading invariant conditions where specification searches were conducted based on 99% confidence interval or the 6.635 cutoff of modification indices*

Sample size	Missing rate	FIML	rFIML	WLSMV_MI	WLSMV_PD
400	Complete	0.040	0.040	**0.020**	0.170
	30%	0.030	0.032	**0.022**	0.150
	50%	**0.032**	0.032	0.038	0.164
1000	Complete	**0.036**	0.040	0.042	0.142
	30%	0.038	0.042	**0.032**	0.164
	50%	**0.030**	0.032	0.034	0.272
2000	Complete	0.030	0.032	**0.022**	0.168
	30%	0.034	0.034	**0.028**	0.216
	50%	**0.024**	0.028	0.050	0.504

Note. Model type I error in loading invariant conditions is defined as the final model obtained from specification searches incorrectly labeled at least one invariant loading as non-invariant. The lowest model-level type I error rate(s) among the methods within a condition are shown in bold. rFIML indicates robust FIML.

Table 11. *Basal model-level type I error rates of methods in loading invariant conditions where specification searches were conducted based on 95 % confidence interval or the 3.841 cutoff of the modification indices*

Sample size	Missing rate	FIML	rFIML	WLSMV_MI	WLSMV_PD
400	Complete	**0.136**	0.144	0.180	0.422
	30%	**0.146**	0.154	0.168	0.362
	50%	**0.146**	0.158	0.180	0.410
1000	Complete	**0.164**	0.168	0.180	0.422
	30%	0.178	0.182	**0.166**	0.426
	50%	**0.148**	0.152	0.184	0.608
2000	Complete	**0.168**	0.174	0.182	0.454
	30%	0.192	0.202	**0.190**	0.506
	50%	**0.170**	0.176	0.196	0.792

Note. The lowest model-level type I error rate(s) within are shown in bold.

By comparing the results in Tables 10 and 11, one can tell that, when specification searches involve multiple parameters testing, the model-level type I error rates was inflated (larger than the type I error rates indicated by the item level cutoff , e.g., 0.05 for 3.841 and 0.01 for 6.636). To control the model-level type I error rate at a certain level, a stricter criterion needs to be used in the specification search.

In general, the performances of $FIML_{mvn}$, robust FIML and WLSMV_MI were comparable. For example, as shown in Table 10, these three methods all kept the model-level type I error rates below 0.05, despite the missing data rates. In contrast, WLSMV_PD always had the highest

model-level type I error rates. Even when 6.635 was used as the cutoff for modification indices, the model-level type I error rates from WLSMV_PD were substantively larger than 0.1 in all conditions (see Table 10). Besides that, except for the conditions with a small sample size (n= 400), the amount of type I error rate inflation of WLSMV_PD increased as the missing data rate increased.

The item-level type I error rates for the loading invariant conditions are presented in Tables 12 and 13. They were calculated as the number of equality constraints that are misidentified as non-invariant within a condition divided by 2500 (500 replications × 5 equality constraints that are examined during the specification searches for loadings. The results presented in these two tables showed the same pattern as those shown in Tables 10 and 11. WLSMV_PD had the highest type I error rates among all methods across conditions. Furthermore, unlike FIML methods and WLSMV_MI, the item level type I error rates of WLSMV_PD increased as the missing data rate increased (except in conditions n = 400)

Table 12. *Basal item level type I error rates of methods in loading invariant conditions where specification searches were conducted based on 99 % confidence interval or the 6.635 cutoff for modification indices*

Sample size	Missing rate	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	0.008	0.008	**0.004**	0.037
	30%	0.006	0.006	**0.004**	0.032
	50%	**0.006**	0.006	0.008	0.034
1000	Complete	**0.007**	0.008	0.010	0.031
	30%	**0.008**	**0.008**	**0.008**	0.036
	50%	**0.006**	**0.006**	0.007	0.061
2000	Complete	0.006	0.007	**0.005**	0.039
	30%	0.007	0.007	**0.006**	0.049
	50%	**0.005**	0.006	0.011	0.132

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. rFIML indicates robust FIML.

Table 13. *Basal item level type I error rates of methods in loading invariant conditions where specification searches were conducted based on 99 % confidence interval or the 3.841 cutoff of modification indices*

Sample size	Missing rate	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	**0.030**	0.031	0.043	0.100
	30%	**0.032**	0.033	0.042	0.091
	50%	**0.032**	0.034	0.047	0.106
1000	Complete	**0.034**	0.035	0.048	0.103
	30%	**0.038**	0.039	0.045	0.108
	50%	**0.031**	0.033	0.048	0.166
2000	Complete	**0.036**	0.038	0.047	0.118
	30%	**0.041**	0.043	0.051	0.126
	50%	**0.038**	0.039	0.052	0.270

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. rFIML indicates robust FIML.

5. 2 Results for threshold invariant conditions

The model-level type I error rates for specification searches for threshold invariant conditions were presented in Table 14. Note that, because FIML and robust FIML do not have threshold parameters, I only presented results from WLSMV_MI and WLSMV_PD. Although the amount of inflation of model-level type I error rates was larger in threshold invariant

conditions than in loading invariant conditions (compare results in Table 14 to the results in Table 10 and Table 11), WLSMV_MI still substantively outperformed WLSMV_PD in controlling the inflations across all conditions, especially when missing data present.

Table 14. *Basal model-level type I error rates of methods in threshold invariant conditions*

Sample size	Missing rate	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
		$\alpha = 0.01$/MFI cutoff: 6.635		$\alpha = 0.05$/MFI cutoff: 3.841	
400	Complete	**0.136**	0.224	**0.392**	0.696
	30%	**0.122**	0.222	**0.429**	0.676
	50%	**0.134**	0.290	**0.452**	0.758
1000	Complete	**0.082**	0.208	**0.338**	0.624
	30%	**0.088**	0.240	**0.350**	0.718
	50%	**0.118**	0.464	**0.376**	0.928
2000	Complete	**0.100**	0.182	**0.356**	0.686
	30%	**0.094**	0.416	**0.362**	0.858
	50%	**0.114**	0.852	**0.388**	0.992

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. $\alpha = 0.01$ and $\alpha = 0.05$ indicate the 99% and 95% confidence intervals used in WLSMV_MI for specification search

The item-level type I error rates for WLSMV_MI and WLSMV_PD in threshold invariant conditions were presented in Table 15. Similar to the results in Table 14, WLSMV_MI can always control the type I error rate at its nominal level, which can be considered as evidence that the substantively-inflated model-level type I error rate in the Table 14 for WLSMV_MI may simply result from the specification searches on thresholds in population modes involving many parameters (23 equality constraints on thresholds in the population models). In fact, the item level type I error rate of the WLSMV_PD approach was also pretty accurate when data are complete; while when missing data present, the item level type I error rate of WLSMV_PD quickly increased (see Table 15).

Table 15. *Basal item level type I error rates of methods in threshold invariant conditions*

Sample size	Missing rate	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
		α = 0.01/MFI cutoff: 6.635		α = 0.05/MFI cutoff: 3.841	
400	Complete	**0.012**	**0.012**	0.059	**0.055**
	30%	**0.011**	0.012	0.058	**0.057**
	50%	**0.011**	0.017	**0.058**	0.074
1000	Complete	**0.008**	0.011	**0.047**	0.050
	30%	**0.008**	0.013	**0.047**	0.064
	50%	**0.009**	0.035	**0.047**	0.127
2000	Complete	**0.009**	**0.009**	**0.048**	0.051
	30%	**0.008**	0.027	**0.046**	0.099
	50%	**0.009**	0.107	**0.049**	0.222

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. α = 0.01 and α = 0.05 indicate the 99% and 95% confidence intervals used in WLSMV_MI for specification search

5.3 Results from loading non-invariant conditions

Perfect recovery indicates that a method can correctly identify all non-invariant parameters without incorrectly labeling any invariant parameters as non-invariant. The results obtained from loading non-invariant conditions with α = 0.01 for specification searches on loading equality constraints are presented in Table 16. I presented the results based on the same condition of α = 0.05 in Appendix C, since the relation between methods did not change when the significance level of α was adjusted. The results related to item level type I error rates and item level power are also presented in Appendix C, given the information they provide are similar to those provided by results related to model-level type I and type II error rates.

In Table 16, all methods share at least two general patterns. First, as the missing data rate increases, the perfect recovery rates in general decrease, except for a few conditions where the perfect recovery rates were always above 0.98 (e.g., conditions with n = 2000). Secondly, when the sample size and amount of non-invariance loading were small, the perfect recovery rates of all methods were low. As for the relative performances between methods, WLSMV_MI has the highest recovery rates in the majority of the conditions. The perfect recovery rates obtained from FIML and robust FIML were similar to each other and usually are not substantively lower than the perfect recovery rates obtained from WLSMV_MI. In fact, when the amount of non-invariance is not small, the perfect recovery rates of FIML were even slightly better than

those of WLSMV_MI. In contrast, the perfect recovery rates for loadings obtained from WLSMV_PD were only better than other methods in two conditions where the amounts of non-invariance are small. Furthermore, its perfect recovery rate can substantively decrease as the missing data increase when the type of non-invariance was set to be non-uniform.

Table 16. *Perfect recovery rates of methods in loading non-invariant condition where specification searches were conducted based on 99 % confidence interval or 6.635 cutoff for modification indices*

Type of DIF	Sample size	Missing rate	FIML	rFIML	WLSMVMI	WLSMVPD
small	400	Complete	0.098	0.104	**0.174**	0.098
		30%	0.060	0.070	**0.116**	0.106
		50%	0.028	0.028	**0.076**	0.088
	1000	Complete	0.430	0.434	**0.570**	0.406
		30%	0.294	0.302	**0.422**	0.388
		50%	0.140	0.148	0.276	**0.454**
	2000	Complete	0.894	0.894	**0.940**	0.784
		30%	0.802	0.802	**0.874**	0.776
		50%	0.598	0.596	0.734	**0.840**
large	400	Complete	0.862	0.858	**0.904**	0.734
		30%	0.710	0.724	**0.780**	0.595
		50%	0.482	0.488	**0.600**	0.493
	1000	Complete	**0.980**	**0.980**	0.974	0.900
		30%	**0.972**	**0.972**	0.970	0.906
		50%	0.926	0.926	**0.946**	0.880
	2000	Complete	0.972	0.972	**0.972**	0.882
		30%	0.972	0.970	**0.978**	0.900
		50%	**0.970**	**0.970**	0.966	0.890
mixed	400	Complete	0.704	0.714	**0.748**	0.654
		30%	0.558	0.564	**0.646**	0.530
		50%	0.380	0.396	**0.512**	0.476
	1000	Complete	0.960	0.958	**0.962**	0.892
		30%	0.898	0.900	**0.920**	0.880
		50%	0.784	0.782	**0.844**	0.838
	2000	Complete	**0.992**	0.990	0.966	0.906
		30%	**0.990**	**0.990**	0.970	0.910
		50%	**0.988**	**0.988**	0.968	0.918
nonuniform	400	Complete	0.860	**0.864**	0.820	0.507
		30%	0.774	0.784	**0.788**	0.188
		50%	0.544	0.556	**0.626**	0.026
	1000	Complete	**0.966**	0.966	0.954	0.772
		30%	**0.970**	0.968	0.952	0.468
		50%	0.940	0.946	**0.948**	0.042
	2000	Complete	**0.982**	**0.982**	0.972	0.866
		30%	**0.984**	0.982	0.974	0.732
		50%	**0.982**	**0.982**	**0.982**	0.076

Note. The highest perfect recovery rate(s) among methods within a condition are shown in bold

Table 17. *Model-level type I error rates in loading non-invariant conditions where specification searches were conducted based on 99 % confidence interval or the 6.635 cutoff for modification indices*

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.052	0.052	**0.034**	0.148
	30%		0.050	0.052	**0.024**	0.140
	50%		0.038	0.040	**0.032**	0.152
1000	Complete		0.050	0.052	**0.028**	0.160
	30%		0.044	0.044	**0.020**	0.164
	50%		0.048	0.048	**0.018**	0.144
2000	Complete		0.030	0.030	**0.022**	0.124
	30%		0.026	0.030	**0.022**	0.128
	50%		0.026	0.026	**0.022**	0.108
400	Complete	large	**0.014**	0.018	0.028	0.163
	30%		0.026	**0.024**	**0.024**	0.165
	50%		0.036	0.036	**0.024**	0.172
1000	Complete		**0.018**	**0.018**	0.026	0.100
	30%		**0.024**	**0.024**	**0.024**	0.090
	50%		**0.022**	**0.022**	0.026	0.098
2000	Complete		**0.028**	**0.028**	**0.028**	0.118
	30%		0.028	0.030	**0.022**	0.100
	50%		**0.030**	**0.030**	0.034	0.110
400	Complete	mixed	0.028	0.028	**0.026**	0.136
	30%		0.026	0.028	**0.022**	0.144
	50%		0.028	0.028	**0.020**	0.133
1000	Complete		**0.018**	0.022	0.022	0.072
	30%		**0.020**	0.024	0.026	0.074
	50%		**0.018**	0.022	0.024	0.080
2000	Complete		**0.008**	0.010	0.034	0.094
	30%		**0.010**	**0.010**	0.030	0.090
	50%		**0.008**	**0.008**	0.030	0.082
400	Complete	nonuniform	**0.024**	**0.024**	0.030	0.174
	30%		0.030	0.030	**0.026**	0.140
	50%		**0.026**	0.030	**0.026**	0.088
1000	Complete		**0.028**	**0.028**	0.038	0.166
	30%		**0.022**	0.024	0.040	0.160
	50%		0.032	**0.028**	0.038	0.132
2000	Complete		**0.018**	**0.018**	0.028	0.134
	30%		**0.016**	0.018	0.026	0.144
	50%		**0.018**	**0.018**	**0.018**	0.108

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. rFIML indicates robust FIML.

The model-level type I error rates in loading non-invariant conditions are presented in Table 17. WLSMV_MI had the lowest model-level type I error rates when the amount of non-invariance is small. $FIML_{mvn}$ and robust FIML tended to have the lowest model-level type I error rates when neither the amount of non-invariance nor the sample size is small. WLSMV_PD generally had the highest model-level type I error rates, except in the conditions where the sample size was large (n = 2000). This tendency to incorrectly misidentify the invariant loadings as non-invariant could be the reason why WLSMV_PD did not perform well in general in terms of perfect recovery rates in loading non-invariant conditions (see Table 16).

The result for model-level type II error rates in loading non-invariant conditions was presented in Table 18. When the type of non-invariance was small and sample size was 400 or 1000, the type II error rates from all methods were high, indicating that none of the methods could effectively identify non-invariant loadings in these conditions. This explains why the perfect recovery rates were also low for these conditions (see Table 16). However, comparing to the other methods, WLSMV_MI had the lowest type II error rates which explains why WLSMV_MI had the highest perfect recovery rates in most loading non-invariant conditions as shown in Table 16.

Table 18. *Model-level type II error rates in loading non-invariant conditions (α = 0.01/cutoff of the modification indices is set at 6.635)*

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.900	0.894	**0.816**	0.892
	30%		0.940	0.930	**0.878**	0.884
	50%		0.972	0.972	0.922	**0.902**
1000	Complete		0.566	0.560	**0.416**	0.550
	30%		0.706	0.698	**0.568**	0.576
	50%		0.860	0.852	0.716	**0.502**
2000	Complete		0.092	0.092	**0.042**	0.140
	30%		0.186	0.182	**0.108**	0.144
	50%		0.392	0.394	0.252	**0.076**
400	Complete	large	0.128	0.128	**0.076**	0.171
	30%		0.276	0.264	**0.202**	0.339
	50%		0.512	0.506	**0.386**	0.458
1000	Complete		0.002	0.002	**0.000**	0.002
	30%		**0.004**	**0.004**	0.008	0.008
	50%		0.054	0.054	**0.030**	0.040
2000	Complete		**0.000**	**0.000**	**0.000**	**0.000**
	30%		**0.000**	**0.000**	**0.000**	**0.000**
	50%		**0.000**	**0.000**	**0.000**	**0.000**
400	Complete	mixed	0.272	0.264	**0.232**	0.267
	30%		0.426	0.420	**0.338**	0.417
	50%		0.612	0.596	0.476	**0.469**
1000	Complete		0.026	0.026	**0.016**	0.042
	30%		0.088	0.082	0.060	**0.056**
	50%		0.204	0.202	0.138	**0.094**
2000	Complete		**0.000**	**0.000**	**0.000**	**0.000**
	30%		**0.000**	**0.000**	**0.000**	**0.000**
	50%		0.004	0.004	0.002	**0.000**
400	Complete	nonuniform	0.124	**0.120**	0.156	0.406
	30%		0.204	**0.194**	**0.194**	0.776
	50%		0.446	0.432	**0.358**	0.974
1000	Complete		**0.006**	**0.006**	0.012	0.094
	30%		**0.010**	0.012	**0.010**	0.482
	50%		0.034	0.032	**0.020**	0.956
2000	Complete		**0.000**	**0.000**	**0.000**	**0.000**
	30%		**0.000**	**0.000**	**0.000**	0.172
	50%		**0.000**	**0.000**	**0.000**	0.918

Note. The lowest model-level type II error rate(s) among methods within a condition are shown in bold. rFIML indicates robust FIML

5.4 Results from threshold non-invariant conditions

The perfect recovery rates and the model-level type I and type II error rates for threshold non-invariant conditions are presented in Tables 19 - 21. Comparing Table 18 to Table 16, it can be found that the perfect recovery rates for non-invariant thresholds were lower than those for non-invariant loadings. When sample size was small (n = 400), most of the perfect recovery rates in the threshold non-invariant conditions were lower than 35%. Furthermore, these perfect recovery rates decreased as missing data rate increased in most conditions, except for a few conditions where the perfect recovery rates were very low (e.g., < 20%) even with complete data.

As for the relative performances of the examined methods, WLSMV_PD produced higher threshold recovery rates than WLSMV_MI when the amount of non-invariance was small or in conditions where the sample size was < 2000, despite the missing data rates. In contrast, WLSMV_MI outperformed WLSMV_PD when sample size was 2000 and the amount of non-invariance was not small, especially when missing data rates were high (e.g., n = 2000 and missing data rate = 50%).

Table 19. *Perfect recovery rate of methods in threshold non-invariant conditions*

Sample size	Missing rate	Type of DIF	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
			α = 0.01/MFI cutoff: 6.635		α = 0.05/MFI cutoff: 3.841	
400	Complete	small	0.016	**0.030**	0.034	**0.060**
	30%		0.016	**0.034**	0.038	**0.084**
	50%		0.006	**0.042**	0.018	**0.070**
1000	Complete		0.072	**0.188**	0.126	**0.168**
	30%		0.054	**0.232**	0.094	**0.148**
	50%		0.030	**0.234**	0.058	**0.074**
2000	Complete		0.246	**0.620**	**0.368**	0.308
	30%		0.182	**0.528**	**0.306**	0.160
	50%		0.108	**0.164**	**0.220**	0.012
400	Complete	large	0.206	**0.518**	**0.322**	0.304
	30%		0.148	**0.430**	0.248	**0.290**
	50%		0.076	**0.338**	0.162	**0.204**
1000	Complete		0.650	**0.806**	**0.586**	0.354
	30%		0.534	**0.724**	**0.546**	0.268
	50%		0.354	**0.490**	**0.422**	0.100
2000	Complete		**0.896**	0.766	**0.630**	0.298
	30%		**0.874**	0.576	**0.620**	0.142
	50%		**0.778**	0.128	**0.598**	0.008
400	Complete	mixed	0.136	**0.346**	0.220	**0.232**
	30%		0.096	**0.322**	0.190	**0.260**
	50%		0.072	**0.254**	0.118	**0.186**
1000	Complete		0.428	**0.692**	**0.486**	0.340
	30%		0.364	**0.628**	**0.436**	0.262
	50%		0.242	**0.456**	**0.334**	0.098
2000	Complete		**0.802**	0.750	**0.656**	0.306
	30%		**0.726**	0.576	**0.618**	0.168
	50%		**0.608**	0.140	**0.584**	0.006
400	Complete	Nonuniform	0.010	**0.388**	0.106	**0.298**
	30%		0.012	**0.114**	0.092	**0.170**
	50%		**0.008**	**0.008**	0.050	**0.024**
1000	Complete		0.256	**0.748**	**0.442**	0.334
	30%		0.186	**0.396**	**0.396**	0.252
	50%		**0.118**	0.026	**0.304**	0.026
2000	Complete		0.766	**0.778**	**0.608**	0.338
	30%		**0.710**	0.520	**0.594**	0.148
	50%		**0.584**	0.024	**0.544**	0.002

Note. The highest perfect recovery rate(s) among methods within a condition are shown in bold. α = 0.01 and α = 0.05 indicate the 99% and 95% confidence intervals used in WLSMV_MI for specification search

The model-level type I error rates in threshold non-invariant conditions are presented in Table 20. WLSMV_MI outperformed WLSMV_PD in all conditions in controlling the model-level type I error rate inflation, especially with the presence of missing data. As shown in

Table 20, missing data did not affect the type I error rates of WLSMV_MI in threshold non-invariant conditions. In contrast, the model-level type I error rates from WLSMV_PD substantively increased as missing data rates increased, especially when sample size was large (up to 89%). This explains why the perfect recovery rates from WLSMV_PD were less satisfactory than those from WLSMV_MI in threshold non-invariance conditions with large sample sizes in Table 19.

Table 20. *Model-level type I error rates of methods in threshold non-invariant conditions*

Sample size	Missing rate	Type of DIF	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
			α = 0.01/cutoff MFI: 6.635		α = 0.05/cutoff MFI: 3.841	
400	Complete	small	**0.108**	0.194	**0.376**	0.654
	30%		**0.106**	0.212	**0.368**	0.640
	50%		**0.120**	0.278	**0.408**	0.734
1000	Complete		**0.132**	0.230	**0.444**	0.686
	30%		**0.136**	0.324	**0.462**	0.780
	50%		**0.120**	0.528	**0.458**	0.908
2000	Complete		**0.094**	0.224	**0.336**	0.678
	30%		**0.104**	0.426	**0.344**	0.836
	50%		**0.108**	0.830	**0.356**	0.988
400	Complete	large	**0.122**	0.218	**0.398**	0.660
	30%		**0.120**	0.250	**0.406**	0.654
	50%		**0.120**	0.290	**0.416**	0.764
1000	Complete		**0.094**	0.186	**0.382**	0.646
	30%		**0.098**	0.266	**0.384**	0.732
	50%		**0.104**	0.502	**0.414**	0.900
2000	Complete		**0.086**	0.234	**0.370**	0.702
	30%		**0.090**	0.424	**0.376**	0.858
	50%		**0.100**	0.872	**0.394**	0.992
400	Complete	mixed	**0.108**	0.232	**0.380**	0.690
	30%		**0.114**	0.236	**0.388**	0.658
	50%		**0.100**	0.274	**0.402**	0.742
1000	Complete		**0.082**	0.232	**0.370**	0.650
	30%		**0.090**	0.310	**0.390**	0.734
	50%		**0.094**	0.504	**0.392**	0.902
2000	Complete		**0.088**	0.250	**0.334**	0.694
	30%		**0.094**	0.424	**0.350**	0.832
	50%		**0.098**	0.860	**0.370**	0.994
400	Complete	nonuniform	**0.102**	0.190	**0.368**	0.620
	30%		**0.100**	0.200	**0.372**	0.616
	50%		**0.104**	0.224	**0.396**	0.702
1000	Complete		**0.108**	0.216	**0.384**	0.666
	30%		**0.134**	0.448	**0.410**	0.886
	50%		**0.108**	0.222	**0.390**	0.662
2000	Complete		**0.120**	0.440	**0.396**	0.852
	30%		**0.118**	0.890	**0.422**	0.994
	50%		**0.134**	0.448	**0.410**	0.886

Note. The lowest model-level type I error rate(s) among methods within a condition are shown in bold. α = 0.01 and α = 0.05 indicate the 99% and 95% confidence intervals used in WLSMV_MI for specification search

The model-level type II error rates in threshold non-invariant conditions are presented in Table 21. The type II error rates were generally low when the amount of non-invariance was small. In fact, for WLSMV_MI, the type II error were in general higher than 20% (i.e., power was lower than 80%), except for few conditions where the amount of non-invariance was large, sample size was medium or large (n =1000 or 2000), and missing data rate was > 30%. These results imply that low perfect recovery rates from WLSMV_MI was likely due to the lack of ability of this method to correctly identify the non-invariant thresholds (see Table 18). On the other hand, the type II error rates from WLSMV_PD were in general lower than those from WLSMV_MI, except in a few conditions where the type of non-invariance was non-uniform and the sample size was large (n = 2000).

Table 21. *Model-level type II error rates of methods in threshold non-invariant conditions*

Sample size	Missing rate	Type of DIF	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
			α = 0.01/cutoff MFI: 6.635		α = 0.05/cutoff MFI: 3.841	
400	Complete	small	0.968	**0.966**	0.878	**0.848**
	30%		0.970	**0.956**	0.878	**0.770**
	50%		0.984	**0.934**	0.932	**0.726**
1000	Complete		0.872	**0.774**	0.696	**0.480**
	30%		0.906	**0.640**	0.732	**0.308**
	50%		0.944	**0.464**	0.812	**0.158**
2000	Complete		0.708	**0.238**	0.462	**0.094**
	30%		0.764	**0.074**	0.522	**0.018**
	50%		0.854	**0.014**	0.642	**0.002**
400	Complete	large	0.736	**0.370**	0.478	**0.156**
	30%		0.812	**0.438**	0.556	**0.178**
	50%		0.882	**0.534**	0.684	**0.168**
1000	Complete		0.318	**0.018**	0.150	**0.008**
	30%		0.430	**0.014**	0.198	**0.002**
	50%		0.620	**0.018**	0.348	**0.000**
2000	Complete		0.036	**0.000**	0.004	**0.000**
	30%		0.066	**0.000**	0.012	**0.000**
	50%		0.176	**0.000**	0.052	**0.000**
400	Complete	mixed	0.816	**0.568**	0.604	**0.296**
	30%		0.862	**0.584**	0.656	**0.282**
	50%		0.902	**0.652**	0.762	**0.262**
1000	Complete		0.530	**0.114**	0.274	**0.052**
	30%		0.588	**0.090**	0.314	**0.024**
	50%		0.710	**0.068**	0.450	**0.010**
2000	Complete		0.136	**0.002**	0.040	**0.000**
	30%		0.214	**0.000**	0.086	**0.000**
	50%		0.336	**0.000**	0.134	**0.000**
400	Complete	nonuniform	0.990	**0.544**	0.856	**0.242**
	30%		0.988	**0.872**	0.876	**0.644**
	50%		**0.990**	**0.990**	**0.930**	0.932
1000	Complete		0.730	**0.066**	0.412	**0.026**
	30%		0.876	**0.956**	**0.586**	0.860
	50%		0.188	**0.000**	0.062	**0.000**
2000	Complete		0.244	**0.136**	0.072	**0.062**
	30%		**0.382**	0.900	**0.140**	0.814
	50%		**0.876**	0.956	**0.586**	0.860

Note. The lowest model-level type II error rate(s) among methods within a condition are shown in bold. α = 0.01 and α = 0.05 indicate the 99% and 95% confidence intervals used in WLSMV_MI for specification search

Chapter 6─Conclusions

In this dissertation, simulation studies have been conducted to examine the effects of ordinal missing data on ME/I tests and specification searches. In this section, I discuss the results of the simulation studies regarding the four research questions I raised in Chapter 2.

6.1 Testing measurement invariance with ordinal missing data

Question 1: How do the different strategies ($FIML_{mvn}$, robust FIML and WLSMV_PD) perform in terms of $\Delta\chi^2$ tests for measurement invariance with the presence of ordinal missing data?

As hypothesized, WLSMV_PD will result in highly inflated type I error rates for the $\Delta\chi^2$ tests in missing data conditions. In contrast, the two FIML-based methods have much better control of type I error rates. Robust FIML slightly outperforms $FIML_{mvn}$ in preventing over-control of the type I error rates in a few conditions, which makes it the best strategy for maintaining type I error rates of $\Delta\chi^2$.

There are two explanations for the inflated type I error rates from WLSMV_PD. First, as mentioned in Chapter 1, WLSMV treats the thresholds and correlation matrix calculated based on the PD as if they are from complete data; thus, it fails to account for the uncertainty due to missing data. Second, PD may result in a non-uniform sample decrease across summary statistics, which distorts the χ^2 test statistic (Bollen, 1989; Kaplan, 2014).

Even though WLSMV_PD tends to have a higher power to detect non-invariance in missing data conditions, this advantage is not really meaningful, given the inflated type I error rates. Both FIML methods have sufficient power to detect non-invariance in loadings and thresholds when the sample sizes or the amounts of non-invariance are moderate or large.

Question 2: How do the different strategies ($FIML_{mvn}$, robust FIML and WLSMV_PD perform in producing accurate parameter estimates and standard error estimates?

WLSMV_PD can still provide accurate point and standard error estimates for loading with missing data. Similar results have been found in previous studies on PD with ML for continuous data (Enders, 2001; Enders & Bandalos, 2001; Savalei & Bentler, 2005).

In contrast, the loading estimates from the FIML methods are only acceptable (mean relative bias < 10%) when thresholds are symmetrically distributed. Similar results have been found in past research on the performance of MLR with ordinal complete data. For example,

Rhemtulla et al. (2012) found that loading estimates from MLR are more accurate when thresholds are symmetric. Last, the standard errors from robust FIML are more accurate than those from $FIML_{mvn}$.

6.2 specification searches with ordinal missing data

Question 3: How do the different strategies ($FIML_{mvn}$, robust FIML and WLSMV_PD) perform in backward MFI search?

As hypothesized, among the three backward specification methods, the MFI with WLSMV_PD tends to misidentify the invariant loadings as non-invariant when missing data present. As a result, WLSMV_PD in general has the worst perfect recovery rates in comparison to the other methods. In contrast, FIML and robust FIML have a better control of type I error rates across all conditions. Both FIML methods tend to perform similarly in terms of type I error rates and perfect recovery rates in most conditions.

Question 4: How does the forward specification search with confidence intervals perform using WLSMV_MI in comparison to the backward search methods using the three strategies in question 3?

My hypothesis that WLSMV_MI is the best specification method is only partially supported. In invariant conditions and loading non-invariant conditions, WLSMV_MI has the lowest type I error rates and the best perfect recovery rates in general. However, in the threshold non-invariant conditions, its perfect recovery rates are worse than those of WLSMV_PD regardless of the missing data rates, unless the sample size is large (n = 2000) and the amount of non-invariance is not small. This suggests that WLSMV_MI requires a large sample size (n = 2000 in this case) to be sufficiently powered to accurately identify non-invariant thresholds.

One possible explanation for the power issue of WLSMV_MI in my thresholds non-invariant conditions could be the low percentage of non-invariant thresholds in population model of my threshold non-invariant conditions (2/23). Forward specification search methods were usually less preferred than the backward specification search methods if the percentage of non-invariant parameters is low (Kim & Yoon, 2014). One of reason is that in such models, forward search methods need to correctly impose many constraints to research a perfect recovery; while the backward methods only need to correctly release few constraints to reach the same purpose. This property gives backward methods an advantage. Future studies can examine this

hypothesis by manipulating the proportion of non-invariant items in the population model.

6.3 Empirical Example

Having examined the performances of the different methods using simulated data, I provided an empirical example to demonstrate their actual data performances. The data for the example was taken from a study on quality of life (QOL) conducted by Chen & Yao (2015). In that study, QOL data were collected from 404 participants in Taiwan using the World Health Organization's Quality of Life Instrument Brief Version (WHOQOL-BREF), which is a self-report scale for QOL with 26 items. Items 1 and 2 are measures of general quality of life and general health status, respectively. The other 24 items are measures of four sub-domains of QOL: Physical health (items 3, 4, 10, 15, 16, 17, and 18), Psychological (items 5, 6, 7, 11, 19, and 26), Social Relationships (items 20, 21, and 22), and Environment (items 8, 9, 12, 13, 14, 23, 24, and 25) (see Yao, 2005*).*

Originally, all of the items in the WHOQOL-BREF were measured on a five-point Likert-type scale. Chen & Yao (2015) aimed to examine whether the reliability and validity of the WHOQOL -BREF could be improved if the format of the original Likert-type scale was revised. Specifically, they compared the psychometric properties of the original WHOQOL-BREF to several novel scales developed based on fuzzy set theory (e.g., Hesketh, Pryor, Gleitzman & Hesketh, 1988). One of the novel scales was a fuzzy scale extended from the visual analogue scale (VAS).

For simplicity, I included the following items in this dissertation: the VAS measure for general QOL (VAS item 1, range of 0–10), six items for the psychological domain in the original WHOQOL-BREF (Likert-type items 5, 6, 7, 11, 19, and 26), and gender (158 males and 240 females). I excluded 6 participants who did not report their genders.

Given that the original data were almost complete (only two participants had missing data on one of the items), I imposed missing data on the last three items in the Psychological domain (i.e., items 11, 19, and 26) for female participants, following the same mechanism used in the simulation study. The missingness was determined by the general QOL measure. I varied the missing data rate at two levels: 30% or 50%. I fit a single-factor model to the six items for the psychology domain, including the general QOL measure as an auxiliary variable. Note that in the dataset containing 50% missing data for female participants, I further collapsed the fourth

category in item 11, due to data sparseness, when conducting analyses with WLSMV_PD and WLSMV_MI. All the empirical ME/I tests and specification searches were conducted in Mplus.8.0 (L. K. Muthén & B. O. Muthén, 1998-2017)

6.3.1 Empirical Example for testing ME/I with ordinal missing data

Similarly to the simulation study 1, empirical ME/I tests were conducted by comparing the scalar invariance models to the configural invariance model using $\Delta\chi^2$ tests. The results were presented in Table 22. When the data were complete, all methods suggested that the factor structure for the psychological domain of WHOQOL-BREF was scalar invariant across genders ($p > .1$). With missing data, the p values of the $\Delta\chi^2$ tests obtained from WLSMV_PD decreased as the proportion of missing data increased. When missing data in the last three items was 50%, WLSMV_PD led to a rejection of the null (invariant) hypothesis ($p < .05$), which was inconsistent with the conclusion based on the complete data and likely a type I error. In contrast, FIML and rFIML were more robust to the presence of ordinal missing data (all p still > .1).

Table 22. *Measurement invariance tests between gender on the psychological domain subscale of the WHOQOL-BREF*

		χ^2	df	$\Delta\chi^2$	*p-value*	CFI	TLI	RMSEA
Complete data								
$FIML_{mvn}$	Configural	46.533	18			.953	.921	.089
	Scalar	57.684	28	11.151	.345	.951	.947	.073
robust FIML	Configural	38.485	18			.955	926	.076
	Scalar	49.846	28	10.481	.399	.952	.949	.063
WLSMV_PD	Configural	56.931	18			.974	.956	.104
	Scalar	64.565	40	21.227	.506	.983	.988	.056
Missing data rate = 30%								
$FIML_{mvn}$	Configural	30.657	18			.975	.958	.059
	Scalar	43.618	28	12.960	.225	.969	.966	.053
robust FIML	Configural	24.899	18			.982	.970	.044
	Scalar	37.262	28	12.211	.271	.976	.974	.041
WLSMV_PD	Configural	33.723	18			.987	.979	.066
	Scalar	59.612	40	29.55	.129	.984	.988	.050
Missing data rate = 50%								
$FIML_{mvn}$	Configural	25.721	18			.983	.972	.046
	Scalar	37.193	28	11.472	.321	.980	.978	.041
robust FIML	Configural	23.214	18			.986	.976	.038
	Scalar	33.406	28	10.215	.421	.985	.984	.031
WLSMV_PD	Configural	33.113	18			.986	.977	.065
	Scalar	63.051	39	32.762	**.048***	.978	.983	.056

6.3.2 Empirical Example for specification searches with ordinal missing data

The same data set was then used to conduct specification searches with $FIML_{mvn}$, robust FIML, WLSMV_PD, and WLSMV_MI. The results were reported in Tables 23 to 25. When conducting specification searches for loadings, I fitted metric invariance models and a configural invariance model to the data using the three backward search methods (FIML, robust FIML, WLSMV_PD) and WLSMV_MI respectively. With WLSMV_MI, the loading of the sixth item is used as the anchor item, given that its modification index is always less than 1.5 across methods in all empirical conditions. In addition, I set the first thresholds within items to be invariant across groups to prevent convergence problems and improper solutions. The invariance constraints are considered reasonable, given that (1) the full scalar invariance model was supported (see Table 22), and (2) the MFI on these constraints were all smaller than 6.635 (i.e., 0.01 critical value for χ^2 statistic with $df = 1$) and all of their corresponding confidence intervals covered zero despite the missing data rates.

The MFI and confidence intervals for non-invariant loading search were presented in Table 23. The MFI obtained from FIML and robust FIML indicated that all loading equality constraints

seemed plausible. The same conclusion was reached if the 99% confidence interval provided by WLSMV_MI was used. In contrast, the MFI provided by WLSMV_PD indicated that the first loading equality should be released (> 6.635) regardless of the missing data rates. The values of MFI on this constraint increased as the missing data rate increased (see Table 23), which is evidence that the MFI from WLSMV_PD are more likely to produce false alarms about non-invariant loadings.

The results for specification searches on non-invariant thresholds are presented in Table 24 and Table 25. The MFI from WLSMV_PD on the first threshold in item 1 are always larger than 3.841 and close the 6.635 cutoff. In contrast, confidence intervals from WLSMV_MI all covered zero, regardless of the missing data rates, suggesting the absence of non-invariant thresholds. Fortunately, it seems that releasing the first threshold in item 1 or not did not have a profound impact on the estimations of latent mean differences across groups of WLSMV_PD in the current example. The latent mean differences between groups were always smaller than 0.1. All the differences were not significant regardless missing data rates ($p > 0.1$), as the results obtained from WLSMV_MI.

Table 23. *Modification indices and 99% confidence intervals for loading equality constraints in metric invariance model*

Missing rate		Modification indices			99% CI
		FIML	rFIML	WLSMVPD	WLSMVMI
Complete	Item 1	0.896	0.752	7.761	(-0.472,0.416)
	Item 2	0.933	0.783	0.260	(-0.37,0.513)
	Item 3	0.184	0.155	0.166	(-0.443,0.356)
	Item 4	1.275	1.070	0.149	(-0.469,0.386)
	Item 5	1.699	1.426	2.205	(-0.628,0.302)
	Item 6	1.302	1.093	0.990	Anchor Item
30%	Item 1	2.157	1.788	8.853	(-0.493,0.31)
	Item 2	1.266	1.050	0.160	(-0.345,0.391)
	Item 3	0.231	0.192	0.375	(-0.437,0.381)
	Item 4	0.791	0.656	0.038	(-0.202,0.628)
	Item 5	3.265	2.706	2.330	(-0.388,0.457)
	Item 6	0.524	0.435	0.719	Anchor Item
50%	Item 1	3.241	2.876	10.395	(-0.369,0.414)
	Item 2	0.820	0.728	1.497	(-0.318,0.433)
	Item 3	0.172	0.153	0.143	(-0.397,0.437)
	Item 4	0.385	0.342	3.032	(-0.177,0.717)
	Item 5	3.261	2.894	1.253	(-0.586,0.381)
	Item 6	0.022	0.020	1.021	Anchor Item

Modification indices larger than 6.635 were highlighted.

Table 24. *Modifications indices on threshold equality constraints obtained from WLSMV_PD in scalar invariance model*

		Threshold 1	Threshold 1	Threshold 3	Threshold 4
Complete	Item 1	4.494	0.520	0.018	0.129
	Item 2	0.001	1.841	0.745	0.152
	Item 3	0.002	3.081	1.098	0.144
	Item 4	0.006	1.511	0.263	0.182
	Item 5	1.342	0.651	1.696	0.113
	Item 6	0.212	0.151	3.325	0.093
30%	Item 1	6.138	2.031	0.092	0.164
	Item 2	0.021	3.121	1.846	0.006
	Item 3	0.072	1.892	1.982	0.029
	Item 4	0.18	2.981	0.133	0.038
	Item 5	0.756	1.842	4.596	0.286
	Item 6	0.032	0.334	0.716	0.113
50%	Item 1	6.274	2.284	0.186	0.087
	Item 2	0.153	4.423	2.363	0.000
	Item 3	0.355	1.220	1.992	0.127
	Item 4	2.464	0.002	0.024	collapsed
	Item 5	0.443	2.443	5.086	0.034
	Item 6	0.183	0.307	0.022	0.232

Table 25. 99% *confidence interval on threshold equality constraints obtained from WLSMV_MI*

		Threshold 1	Threshold 1	Threshold 3	Threshold 4
Complete	Item 1	(-1.401,1.374)	(-1.454,0.533)	(-1.44,0.497)	(-1.503,0.718)
	Item 2	(-2.301,0.366)	(-1.654,0.526)	(-1.521,0.607)	(-1.469,0.733)
	Item 3	(-1.737,0.205)	(-1.608,0.090)	(-1.047,0.688)	(-1.064,0.931)
	Item 4	(-2.450,0.369)	(-1.994,0.071)	(-1.424,0.559)	(-1.131,1.004)
	Item 5	(-2.905,0.154)	(-2.066,0.370)	(-1.936,0.443)	(-1.642,0.771)
	Item 6	Anchor Item	(-0.451,1.284)	(-0.591,1.222)	(-0.416,1.675)
30%	Item 1	(-1.318,1.461)	(-1.373,0.623)	(-1.358,0.585)	(-1.422,0.808)
	Item 2	(-2.194,0.457)	(-1.550,0.620)	(-1.418,0.702)	(-1.365,0.827)
	Item 3	(-1.661,0.277)	(-1.531,0.162)	(-0.968,0.757)	(-0.989,1.004)
	Item 4	(-2.166,0.721)	(-2.121,0.077)	(-1.34,0.716)	(-1.065,1.083)
	Item 5	(-2.378,0.586)	(-1.963,0.466)	(-1.862,0.469)	(-1.597,0.771)
	Item 6	Anchor Item	(-0.495,1.235)	(-0.590,1.259)	(-0.430,1.872)
50%	Item 1	(-1.431,1.470)	(-1.509,0.655)	(-1.498,0.622)	(-1.554,0.836)
	Item 2	(-2.269,0.461)	(-1.636,0.635)	(-1.505,0.718)	(-1.455,0.846)
	Item 3	(-1.756,0.294)	(-1.634,0.186)	(-1.070,0.781)	(-1.078,1.015)
	Item 4	(-2.122,0.316)	(-1.403,0.846)	(-1.158,1.186)	collapsed
	Item 5	(-2.494,0.796)	(-2.078,0.629)	(-1.928,0.580)	(-1.615,0.873)
	Item 6	Anchor Item	(-0.651,1.124)	(-0.814,1.090)	(-0.962,1.418)

6.4 Conclusions

In this dissertation, I have compared the relative performances of different methods for testing ME/I and conducting specification searches with simulated and empirical five-point Likert-type data. The results indicate that when ordinal missing data present, WLSMV_PD should be avoided for ME/I testing and specification searches, given the potential type I error rate inflation. In contrast, robust FIML can be a reasonable choice for ME/I testing, for it provides the best control of type I error rates among the examined methods and has enough power to detect non-invariant items when the amount of non-invariance is not small. Note that a lack of power to detect small amounts of non-invariance (e.g., 0.2 on standardized loadings) is still a limitation, though researchers have found that in many cases, this limitation does not significantly affect follow-up analysis such as cross-groups (cross-populations) comparisons for slopes between latent factors and factor score estimations (e.g., Curran, Cole, Bauer, Hussong, & Gottfredson, 2016; Shi, Song & Lewis, 2017).

As for the specification searches, none of the methods examined in the study were superior under all conditions. Moreover, all methods had the model inflated level type I error rates. Although WLSMV_MI was a relatively effective method in the invariant and loading non-invariant conditions, it still failed to correctly detect non-invariant thresholds (perfect

recovery rate > 0.8) unless the sample size was 2000, which is larger than the sample sizes in most psychological studies. This property could be a common limitation shared by all traditional specification searches based on confidence intervals and modification indices (i.e., the lack of ability to recover partial invariance models when the number of parameters that need to be searched is large). More detailed discussions about these issues (and their potential methodological solutions) are presented in section 6.6; in this section, however, I focus my discussion on two strategies that researchers might use to enhance the power of locating non-invariant thresholds within the simulation conditions I used in this study. One possible strategy that researchers could use is to identify more anchor items to reduce the number of equality constraints that need to be examined before the search. In the threshold non-invariant conditions in study 2, I only set the first threshold of item 1 as the anchor item (parameter) and made the other 23 thresholds the targets for specification searches. It is conservative to assume that researchers will know only one anchor (invariant) parameter. Recently, methodologists have proposed several effective methods that can help researchers to effectively find anchor (invariant) items before searching in frequentist and Bayesian frameworks (e.g., Jung and Yoon, 2017; Shi, Song, Liao, Terry & Snyder, 2017). As the number of anchor items increases, the number of threshold constraints that need to be searched will decrease, which could increase the perfect recovery rates and the model-level power to locate non-invariant thresholds.

Another strategy empirical researchers may use is to change the goal of specification searches from correctly identifying any non-invariant threshold to finding any item that contains one or more non-invariant thresholds. The former goal will be appropriate if the goal is to accurately release the constraints that involve non-invariant parameters to preserve the model's simplicity. The latter goal will be appropriate if one intends to discard or revise the items that involve non-invariant parameters. In this case, one only needs to correctly identify which "items" contain non-invariant thresholds without the need to know which thresholds are non-invariant within an item. As a result, the probability of successfully conducting specification searches in threshold non-invariant conditions would increase. In addition, if researchers change the goal of search to only find out which "items" have non-invariant thresholds, then intercept invariance testing of the two FIML based methods might also be used as a proxy to tell researchers whether an item has any non-invariant thresholds. Future studies are needed to examine the efficacy of

this approach.

6.5 Suggestions for empirical researchers

A major contribution of this dissertation is to demonstrate the influences of ordinal missing data when conducting ME/I tests and specification search with SEM. An important result of the current study that I consistently found across the conditions is that the default pairwise deletion method for WLSMV could cause severe type I error rate inflation when missing data present. Thus, I recommend that researchers avoid the $\Delta\chi^2$ and the modification indices of WLSMV_PD as the only criterion for ME/I testing and specification search if there is a substantive amount of missing data in the dataset. Second, empirical researchers should be aware that when specification searches are conducted, stricter type I error rates should be used for implementing item level search to mitigate the inflation of the type I error rates at the model level.

These suggestions could be useful for clinical psychologists working on clinical samples or developmental psychologists working on longitudinal studies, given that such studies are likely to have a substantive amount of missing data. In addition, these suggestions should also be useful for researchers who want to report the covariance and correlation matrix in their SEM studies. FIML-based methods are the default setting for many software programs when data are treated as continuous. However, if researchers just use the traditional procedures to calculate these summary statistics (e.g., calculate the covariance between variables pair-wisely), it is likely that the default listwise or pairwise deletion will be used during the calculations without being noticed. These properties could directly hinder the abilities of other researchers to reproduce the results-based covariance matrices provided based on raw data set that contains missing data and lead to inaccurate conclusions, such as failure to reproduce the results of previous studies.

6.6 Limitations

Similar to other studies, this study has several limitations that should be noted when interpreting the results I obtained from this dissertation. First, in this dissertation, I did not manipulate the number of categories within an item but always used five-point items for all conditions. Previous studies have shown that the number of categories within an item can change the performances of continuous estimators (e.g., Rhemtulla et al., 2012). Thus, it will be interesting and important to investigate whether robust FIML can maintain both its ability to

control type I error rates of $\Delta\chi^2$ tests and its ability to recover the partial metric invariance models when the number of categories within an item is less than five (e.g., dichotomous items).

Second, in this dissertation, I did not examine whether it is appropriate to use "confidence intervals" obtained from FIML, robust FIML, or WLSMV_PD for specification search. Specifically, in the current study, I only compared the specification searches of these three methods that base their "modification indices" to the search of WLSMV_MI based on its confidence interval. With this research design, it is confounded whether the differences between these methods are caused by the missing data methods (e.g., multiple imputations vs. pairwise deletion) or by the statistics used (confidence intervals vs. modification indices). Instead, I can only propose which strategy is relatively better overall. Future studies could further clarify this issue by including the specification search based on the confidence interval obtained from these methods in simulations.

In addition to the above limitations, in study 2, I also did not manipulate the model complexity or proportions of the non-invariant parameters in the population models when comparing search methods. Instead, in both thresholds and loading non-invariant conditions, I always used a one factor, six-item population model and set two parameters (either loadings or thresholds) to be non-invariant. This setting not only limits the generalizability of my results (e.g., I cannot examine whether the model level type I error rate will be further inflated as the model becomes larger or more complicated), but also hinders my ability to systematically understand the cause(s) of differences between methods and conditions. As for the proportions of the non-invariant parameters in the populations, given that I always fixed the number of non-invariant parameters as two, the proportion of the non-invariant items in the population model were actually different in my loading non-invariant models and thresholds non-invariant models (2 of 5 loadings vs. 2 of 23 thresholds). This difference makes it difficult for me to compare the performance of method cross loading and threshold non-invariant conditions. For example, with the current research design, I cannot tell which reason (change in target parameters or change in the proportion of non-invariant items) is the deciding factor that makes perfect recovery rates of WLSMV_MI generally better in loading non-invariant conditions than those in thresholds non-invariant conditions.

Fifth, the software I used for ME/I testing and specification searches could also be another

source of confound. In this dissertation, I used Mplus for data analysis. Other SEM software packages, such as LISEREL or lavaan, use different methods to estimate models with ordinal data. Future studies might also seek to further clarify the influences of software packages. Finally, I assumed that latent factors were normally distributed in the simulation studies, which may not necessarily be true in practice. Suh (2015) found that non-normally distributed latent variables can affect the performance of $\Delta\chi^2$ tests obtained from WLSMV and ordinal ML in the context of ME/I testing. Furthermore, non-normality can confound the missing data mechanism that researchers use to generate missing data. For example, in the current study, I simulated missing data such that missing data were more likely to occur with higher auxiliary scores. Graham (2003) considered this way of generating missing data as a linear MAR. It is also possible to generate missing data on both sides of the variables and create non-linear (or convex) MAR (e.g., imposing missingness on participants with high and low auxiliary scores). Graham found that when data are normally distributed, FIML worked well for either linear or non-linear MAR. However, Savalei and Falk (2014) showed that when data were non-normally distributed, the performance of robust FIML can be very sensitive to the different kinds of MAR mechanisms. Future research could further investigate the joint effects of non-normally distributed data and different missing data mechanisms on ME/I testing.

6.7 Possible future directions

The above limitations aside, future studies are needed to examine some other interesting topics in the areas of ordinal missing data issues (and ME/I testing). First, in this dissertation, I only examined some estimators that are widely available across SEM software (e.g., WLSMV, ML, and MLR); however, many other estimators could also have the potential to address the ordinal missing data issues in ME/I testing. Second, in this dissertation, I only used the MG-CFA as the model for ME/I testing, but many models have been proposed to test ME/I with more complicated data sets, which also create different kinds of issues related to ordinal missing data. In the following paragraphs, I briefly discuss the possible research directions from these two perspectives.

As for using different estimators to conducting ME/I testing or specification search with ordinal missing data, I think the Bayesian estimator could be a promising option. Similar to modern missing data techniques, the Bayesian estimator can appropriately address data that are

missing at random. Furthermore, like multiple imputation, it has the flexibility to handle ordinal missing data (e.g., B. O. Muthén, L. K. Muthén, & Asparouhov, 2015). Researchers have proposed several methods to test (approximate) ME/I or to conduct specification search in a Bayesian framework. For example, Shi, Song, Liao, Terry & Synder (2017) proposed that researchers could first use the posterior mean of a selection index that contains the information of differences in a paired of item (standardized differences in loadings and intercepts of an item across groups) and an informative prior (zero mean and small variance) to select an anchor item. After that, the specification search could be conducted by referring to the credible intervals of differences between corresponding loadings and intercepts with a non-informative prior. B.O. Muthén & Asparouhov (2013) also proposed a two-stage procedure of building a partial invariance model with informative priors and credible intervals directly, without searching to locate anchor items first. Even though these two studies still assume that researchers have continuous complete data, given the flexibility of Bayesian modeling for missing data, I think it will be interesting to see how these methods perform when ordinal missing data present.

The other estimator that I believe has the potential to handle the ordinal missing data problem in ME/I testing or specification search is the SEM model with penalized likelihood. For example, Huang (2018) proposed a method to implement the penalized likelihood method to a multiple-groups SEM model. In that article, Huang first decomposed each group parameter into two parts: a common reference component and a group-specific component. After that, by penalizing the group-specific components during the estimations, the non-invariant (heterogeneous) parameter could be found, given that the non-substantively (null) group specific should diminish. An advantage of the penalized approach is that researchers do not have to locate non-invariant parameters by fitting partial invariance models sequentially, which could become cumbersome with complicated models having many parameters. Again, even though Huang's method was proposed based on the assumption that researchers have complete continuous data, the package he developed for penalized SEM already has the ability to use FIML with auxiliary variables. Thus, it will be interesting to compare his approach to the two FIML methods I included in the current dissertation.

In addition to examining the performances of other estimators, the other possible direction in which I believe future research is needed is examination of the issue of ordinal missing data in

other population models, models that are more complicated/flexible than MG-CFA. For example, recently Bauer (2017) proposed to test ME/I in a moderated non-linear factor model (MNLFA). In comparison to MG-CFA or the traditional multiple indicator multiple cause model, MNLFA is a more flexible framework for ME/I testing. It allows researchers to test ME/I across multiple categorical and continuous groupings simultaneously by allowing the parameters in the target model to be specified as the function of grouping variables. This approach not only increases the type of grouping variables that researchers could use for ME/I testing but could also increase the possibility of having missing data (as more groups are involved for ME/I testing). I think further research is needed to examine the influence of ordinal missing data on ME/I testing in this new framework.

To sum up, in the current dissertation, I have examined the influences of ordinal missing data on ME/I testing and specification search. I found that when missing data present, researchers should avoid using WLSMV_PD for ME/I testing specification searches. The ME/I testing methods with modern missing data techniques could effectively control the type I error rates of $\Delta\chi^2$ tests. They are also relatively better in controlling the type I error rates and the inflation of the type I error rates of specification search at the model level, but the fact that WLSMV_MI is lacking in its ability to detect non-invariant thresholds (or the ability to identify the non-invariant parameters in a relatively large model) is an issue that deserves more attention in future studies.

Reference

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling: Issues and techniques*,*243*, 277.

Asparouhov, T (2017, Feb 21). Multiple imputation. [Online forum comment]. Message posted http://www.statmodel.com/cgi-bin/discus/show.cgi?22/381

Asparouhov, T. & Muthén, B. O. (2006). Robust chi-square difference testing with mean and variance adjusted test statistics. Retrieved from https://www.statmodel.com/download/MI7.pdf

Asparouhov, T., & Muthén, B. O. (2008). Chi-square statistics with multiple imputation. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.566.2877&rep=rep1&type=pdf

Asparouhov, T., & Muthén, B. O. (2010)a. Simple second order chi-square correction

Asparouhov, T., & Muthén, B. O. (2010)b. Bayesian analysis using Mplus. Retrieved from https://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. O. (2010)c. Multiple imputation with Mplus. Retrieved from https://www.statmodel.com/download/Imputations7.pdf

Asparouhov, T., & Muthén, B. O. (2010)d. Weight least squared estimation with missing data. Retrieved from https://www.statmodel.com/download/ GstrucMissingRevision.pdf

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. *Handbook of structural equation modeling*, 495-511.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62-83

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social*

Psychology, 95(5), 1005

Chen, P. Y., & Yao, G. (2015). Measuring quality of life with fuzzy numbers: in the perspectives of reliability, validity, measurement invariance, and feasibility. *Quality of Life Research*, 24(4), 781-785.

Chou, C. P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*,25(1), 115-136.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling*, 23(6), 827-844.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, 21(3), 425-438.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4), 352

Enders, C. K. (2010).*Applied missing data analysis*. Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430-457.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, 14(3), 275.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, 25(2), 520.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10(1), 80-100.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, 5(980), 1-16

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12130

Jia, F. (2016). *Methods for handling non-normal missing data in structural equation modeling*. Ph.D.dissertation, University of Kansas, United States

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: forward, backward, and factor-ratio tests. *Structural Equation Modeling*, *23*(4), 567-584.

Jung, E., & Yoon, M. (2017). Two-Step Approach to Partial Factorial Invariance: Selecting a Reference Variable and Identifying the Source of Noninvariance. *Structural Equation Modeling*, 24(1), 65-79.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Guilford Publications.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18(2), 212-228.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, 11(4), 514-534

Li, C. H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 1-14.

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, 21(3), 369.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models:

Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling*, 5(1), 22-36

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling*, *24*(3), 383-394.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9(1), 93.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115-132.

Muthén, B. O. (1998-2004) Mplus technical appendices. Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.stamodel.com/downlaad/techappen.pdf

Muthén, B. O. (2017, Feb 26). Multiple imputation [Online forum comment]. Message posted www.statmodel.com/discussion/ messages/22 /381.html?1488130113

Muthén, L. K. (2010, Sep 14). Model constraints. [Online forum comment]. Message posted www.statmodel.com/discussion/messages/11/535.html?1469624214

Muthén, L. K (2011, Jul, 26) Modification indices. [Online forum comment]. Message posted http://www.statmodel.com/discussion/messages/ 9/153.html?1457176945

Muthén, B. O. & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web. Notes: No. 17. January 11, 2013. Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eight Edition. Los Angeles, CA: Muthén & Muthén

Muthén, B. O, Muthén, L. K & Asparouhov, T. (2015). Estimator choices with categorical outcomes. Mplus Web Notes: March 2015. Retrieved from https://www.statmodel.com/download/EstimatorChoices.pdf

Muthén, B. O. & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model.

Psychometrika, 60(4), 489-503.

Muthén, B. O, du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes Unpublished Technical Report.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*,17(3), 354.

Rosseel, Y. (2012).lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36

Rosseel, Y. (2016 Sep). lavaan version history. Retrieved from http://lavaan.ugent.be/history/dot5.html

Rosseel, Y. (2017, Feb 12) Missing data and WLSMV. Retrieved from https://groups.google.com/forum/#!topic/lavaan/J1TtmcPDTb0

Rubin, R. B. (1987). *Multiple imputation for nonresponse in survey*s J Wiley & Sons, New York, NY.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, 21(2), 167-180.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C.Clogg (Eds.), *Analysis of latent variables in developmental research* (pp. 399-419).Newbury Park, CA: Sage

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507-514

Savalei, V. (2008). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling*, 15(1), 1-22.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21(1), 149-160.

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete

nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling*, 12(2), 183-214.

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling*, 21(2), 280-302.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.

Shi, D., Song, H., & Lewis, M. D. (2017). The Impact of Partial Factorial Invariance on Cross-Group Comparisons. *Assessment*, DOI: 1073191117711020.

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Rresearch*, 52(4), 430-444.

Suh, Y. (2015). The Performance of Maximum Likelihood and Weighted Least Square Mean and Variance Adjusted Estimators in Testing Differential Item Functioning With Nonnormal Trait Distributions. *Structural Equation Modeling*, 22(4), 568-580.

Teman, E.D. (2012). *The performance of multiple imputation and full information maximum likelihood for missing ordinal data in structural equation models*. Ph.D.dissertation, University of North Colorado, United States.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.

Widaman, K. F., Grimm, K. J., Early, D. R., Robins, R. W., & Conger, R. D. (2013). Investigating factorial invariance of latent variables across populations when manifest variables are missing completely. *Structural Equation Modeling*, 20(3), 384-408

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 281-324.

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research*, 50(5), 484-503.

Xu, Y & Green, S.B. (2015). The impact of varying the number of measurement invariance constraints on the assessment of between group differences of latent means. *Structural Equation Modeling*, 00, 1-12.

Yao, G. (2005). *Development and manual of the WHOQOL-BREF Taiwan version* (2nd ed). Taipei, Taiwan. The WHOQOL Group

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, 46(4), 1199-1206.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, 14(3), 435-463.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30(1), 165-200.

Yuan, K. H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods & Research*, 34(2), 240-258.

Zhou, A.Q., Whealin, J.M., Wang, C. & Lee, R.M. (2017) A measure of perceived family stigma: Validity in a military sample. *Psychological Assessment*, 29(9), 1167.

Appendix A

In appendix A, I present five Mplus syntaxes. Appendices A1 and A2 were the syntaxes of ME/I tests with FIML and WLSMV_PD when ordinal missing data present. On the other hand, the syntax in Appendix A3 was used for backward MFI specification search. Last, syntax A4 to A6 were the Mplus syntaxes that will be used in forward CI specification search with multiple imputation.

Appendix A1

Mplus syntax of testing ME/I with FIML

```
TITLE:
      ML invariance with FIML data;
DATA:
     file is exampleDat.txt;
VARIABLE:
      names are
         group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 aux1;
      usevariables are
         group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10;
   auxiliary = (m) aux1;
         GROUPING = group (0 = a 1 = b);
         missing are all (999);
   MODEL: f1 by v1* v2 v3 v4 v5 v6 v7 v8 v9 v10;
   f1@1
      ANALYSIS:
   MODEL = CONFIGURAL SCALAR;
      ESTIMATOR = ML;
```

Appendix A2

```
TITLE:

   WLSMVPD with manual configural invariance;
   DATA:
   file is example.txt;
   VARIABLE:
   names are
      group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 aux1;

   usevariables are group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 aux1;
      categorical are v1 v2 v3 v4 v5 v6 v7 v8 v9 v10;
      GROUPING = group (0 = a 1 = b);
      missing are all (999);


MODEL: f1 by v1* v2 v3 v4 v5 v6 v7 v8 v9 v10;

f1@1;
aux1 with v1-v10;
{v1-v10@1};

model b:
f1 by v1* v2 v3 v4 v5 v6 v7 v8 v9 v10;
f1@1;
[f1@0];

[v1$1];
[v1$2];
[v1$3];
[v1$4];


[v2$1];
[v2$2];
[v2$3];
[v2$4];


[v3$1];
[v3$2];
[v3$3];
[v3$4];
```

```
[v4$1];
[v4$2];
[v4$3];
[v4$4];


[v5$1];
[v5$2];
[v5$3];
[v5$4];


[v6$1];
[v6$2];
[v6$3];
[v6$4];


[v7$1];
[v7$2];
[v7$3];
[v7$4];
{v7@1};

[v8$1];
[v8$2];
[v8$3];
[v8$4];


[v9$1];
[v9$2];
[v9$3];
[v9$4];


[v10$1];
[v10$2];
[v10$3];
[v10$4];

ANALYSIS:
ESTIMATOR = WLSMV;
```

SAVEDATA: DIFFTEST=manualConf.dif;

```
TITLE:

   WLSMV manual scalar invariance;

   DATA:

   file is example.txt;

   VARIABLE:

   names are
      group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 aux1;

   usevariables are group v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 aux1;
      categorical are v1 v2 v3 v4 v5 v6 v7 v8 v9 v10;
      GROUPING = group (0 = a 1 = b);
      missing are all (999);

MODEL: f1 by    v1* (L1)
                    v2    (L2)
                    v3    (L3)
                    v4    (L4)
                    v5    (L5)
                    v6    (L6)
                    v7    (L7)
                    v8    (L8)
                    v9    (L9)
                    v10 (L10);

      f1@1;
      aux1 with v1-v10;

{v1-v10@1};

model b:

f1 by    v1* (L1)
                    v2    (L2)
                    v3    (L3)
                    v4    (L4)
                    v5    (L5)
                    v6    (L6)
                    v7    (L7)
                    v8    (L8)
                    v9    (L9)
```

```
                    v10 (L10);
{v1-v10};
f1;
   ANALYSIS:
DIFFTEST = manualConf.dif;
   ESTIMATOR = WLSMV;
```

Appendix A3
Mplus syntax of backward MFI search on non-invariant loadings with FIML

```
TITLE:
   FIML scalar invariance with modification indices;
DATA:
   file is SixItemDat.txt;
VARIABLE:
   names are
      group v1 v2 v3 v4 v5 v6 aux1;
   usevariables are
      group v1 v2 v3 v4 v5 v6 aux1;
      !categorical are v1 v2 v3 v4 v5 v6 v7 v8 v9 v10;
      GROUPING = group (0 = a 1 = b);
      missing are all (999);
MODEL: f1 by    v1* (L1)
                v2    (L2)
                v3    (L3)
                v4    (L4)
                v5    (L5)
                v6    (L6);
            f1@1;
      aux1 with v1-v6;
model b:
f1;
[f1];
ANALYSIS:
ESTIMATOR = ML;
OUTPUT: MODINDICES (6.635);
```

Appendix A4
Multiple imputation

```
TITLE:
   multiple imputation for group A data;
data:
file is mplusDatA.txt;
variable:
names are group V1 V2 V3 V4 V5 V6 aux;
usevariables are
  V1 V2 V3 V4 V5 V6 aux;
missing are all (999);
   DATA IMPUTATION:
   ! INCOMPLETE VARIABLES TO BE IMPUTED;
   ! (C) FOLLOWING A VARIABLE NAME DENOTES A CATEGORICAL VARIABLE;
     ndatasets = 20;
     IMPUTE = V2 (c) V4 (c);
     save = groupAimp*.dat;

   ANALYSIS:
   process=7;
   type = basic;
   OUTPUT: TECH8;
```

Appendix A5
loading specification search

```
TITLE:

   code for foward CI search on DIF loadings with MI;

   DATA:

   file is mergedDatlist.dat;
type = imputation;

   VARIABLE:

   names are
      group v1 v2 v3 v4 v5 v6 aux1;

   usevariables are
      group v1 v2 v3 v4 v5 v6;
      categorical are v1 v2 v3 v4 v5 v6;
      auxiliary = aux1;
      GROUPING = group (0 = a 1 = b);
      missing are all (999);

!data imputation:
!IMPUTE = v4 (c) v5 (c) v6 (c);
!NDATASETS = 20;

MODEL:
f1 by v1@1; ! anchor item
f1 by v2    (L21);
f1 by v3    (L31);
f1 by v4    (L41);
f1 by v5    (L51);
f1 by v6    (L61);


[V1$1] (T11); ! set to be invariant across groups for idenfication purpose
[V1$2] (T121);
[V1$3] (T131);
[V1$4] (T141);

[V2$1] (T211);
[V2$2] (T221);
[V2$3] (T231);
```

```
[V2$4] (T241);

[V3$1] (T311);
[V3$2] (T321);
[V3$3] (T331);
[V3$4] (T341);

[V4$1] (T411);
[V4$2] (T421);
[V4$3] (T431);
[V4$4] (T441);

[V5$1] (T511);
[V5$2] (T521);
[V5$3] (T531);
[V5$4] (T541);

[V6$1] (T611);
[V6$2] (T621);
[V6$3] (T631);
[V6$4] (T641);

{V1@1};
{V2@1};
{V3@1};
{V4@1};
{V5@1};
{V6@1};


model b:
f1 by v1@1;    ! anchor item
f1 by v2    (L22);
f1 by v3    (L32);
f1 by v4    (L42);
f1 by v5    (L52);
f1 by v6    (L62);

[V1$1] (T11);      ! set to be invariant across groups for idenfication purpose
[V1$2] (T122);
[V1$3] (T132);
[V1$4] (T142);

[V2$1] (T212);
[V2$2] (T222);
```

```
[V2$3] (T232);
[V2$4] (T242);

[V3$1] (T312);
[V3$2] (T322);
[V3$3] (T332);
[V3$4] (T342);

[V4$1] (T412);
[V4$2] (T422);
[V4$3] (T432);
[V4$4] (T442);

[V5$1] (T512);
[V5$2] (T522);
[V5$3] (T532);
[V5$4] (T542);

[V6$1] (T612);
[V6$2] (T622);
[V6$3] (T632);
[V6$4] (T642);

model constraint:
new(ITL2);
new(ITL3);
new(ITL4);
new(ITL5);
new(ITL6);

ITL2 = L21 -L22;
ITL3 = L31 -L32;
ITL4 = L41 -L42;
ITL5 = L51 -L52;
ITL6 = L61 -L62;
analysis:
   process=7;
estimator= WLSMV;
OUTPUT: Cinterval;
```

Appendix A6
Threshold specification search

```
TITLE:
   Eample code in dissertation proposal of foward CI search on DIF thresholds;
   DATA:
   file is mergedDatlist.dat;
type = imputation;
   VARIABLE:
   names are
      group v1 v2 v3 v4 v5 v6 aux1;
   usevariables are
      group v1 v2 v3 v4 v5 v6;
      categorical are v1 v2 v3 v4 v5 v6;
      auxiliary = aux1;
      GROUPING = group (0 = a 1 = b);
      missing are all (999);

!data imputation:
!IMPUTE = v4 (c) v5 (c) v6 (c);
!NDATASETS = 20;
MODEL:
f1 by v1@1;
f1 by v2;
f1 by v3;
f1 by v4;
f1 by v5;
f1 by v6;
[V1$1] (T11); ! set to be invariant across groups for idenfication purpose
[V1$2] (T121);
[V1$3] (T131);
[V1$4] (T141);

[V2$1] (T211);
[V2$2] (T221);
[V2$3] (T231);
[V2$4] (T241);

[V3$1] (T311);
[V3$2] (T321);
[V3$3] (T331);
[V3$4] (T341);
[V4$1] (T411);
[V4$2] (T421);
[V4$3] (T431);
```

```
[V4$4] (T441);
[V5$1] (T511);
[V5$2] (T521);
[V5$3] (T531);
[V5$4] (T541);
[V6$1] (T611);
[V6$2] (T621);
[V6$3] (T631);
[V6$4] (T641);

!f1@1;
{V1@1};
{V2@1};
{V3@1};
{V4@1};
{V5@1};
{V6@1};
model b:
f1;
[f1];
[V1$1] (T11);      ! set to be invariant across groups for idenfication purpose
[V1$2] (T122);
[V1$3] (T132);
[V1$4] (T142);

[V2$1] (T212);
[V2$2] (T222);
[V2$3] (T232);
[V2$4] (T242);

[V3$1] (T312);
[V3$2] (T322);
[V3$3] (T332);
[V3$4] (T342);

[V4$1] (T412);
[V4$2] (T422);
[V4$3] (T432);
[V4$4] (T442);

[V5$1] (T512);
[V5$2] (T522);
[V5$3] (T532);
[V5$4] (T542);
```

```
[V6$1] (T612);
[V6$2] (T622);
[V6$3] (T632);
[V6$4] (T642);

model constraint:
new(IT12);
new(IT13);
new(IT14);

new(IT21);
new(IT22);
new(IT23);
new(IT24);

new(IT31);
new(IT32);
new(IT33);
new(IT34);

new(IT41);
new(IT42);
new(IT43);
new(IT44);

new(IT51);
new(IT52);
new(IT53);
new(IT54);

new(IT61);
new(IT62);
new(IT63);
new(IT64);

IT12 = T122- T121;
IT13 = T132- T131;
IT14 = T142- T141;

IT21 = T212- T211;
IT22 = T222- T221;
IT23 = T232- T231;
IT24 = T242- T241;
```

```
IT31 = T312- T311;
IT32 = T322- T321;
IT33 = T332- T331;
IT34 = T342- T341;

IT41 = T412- T411;
IT42 = T422- T421;
IT43 = T432- T431;
IT44 = T442- T441;

IT51 = T512- T511;
IT52 = T522- T521;
IT53 = T532- T531;
IT54 = T542- T541;

IT61 = T612- T611;
IT62 = T622- T621;
IT63 = T632- T631;
IT64 = T642- T641;

analysis:
   ESTIMATOR = WLSMV;
      process=7;
OUTPUT: Cinterval;
```

Appendix B

Table B1

Mean relative bias of loading estimates across items in group A with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	0.057	0.057	0.057	0.057	0.057	0.057	0.058	0.058	0.058
	0	0.2	0.057	0.057	0.057	0.053	0.053	0.053	0.055	0.055	0.055
	0	0.3	0.058	0.058	0.058	0.059	0.059	0.059	0.057	0.057	0.057
	0	0.4	0.053	0.053	0.053	0.058	0.058	0.058	0.057	0.057	0.057
	0	0.5	0.047	0.047	0.047	0.054	0.054	0.054	0.058	0.058	0.058
	0.2	0	0.054	0.054	0.054	0.058	0.058	0.058	0.056	0.056	0.056
	0.3	0	0.054	0.054	0.054	0.053	0.053	0.053	0.055	0.055	0.055
	0.4	0	0.048	0.048	0.048	0.058	0.058	0.058	0.055	0.055	0.055
	0.5	0	0.054	0.054	0.054	0.054	0.054	0.054	0.057	0.057	0.057
WLSMVPD	0	0	0.008	0.008	0.008	0.003	0.003	0.003	0.002	0.002	0.002
	0	0.2	0.008	0.008	0.007	0.001	0.001	0.001	0.000	0.000	0.000
	0	0.3	0.009	0.009	0.009	0.004	0.004	0.004	0.001	0.001	0.001
	0	0.4	0.004	0.004	0.004	0.004	0.004	0.004	0.002	0.002	0.002
	0	0.5	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.003	0.003
	0.2	0	0.006	0.006	0.006	0.004	0.004	0.004	0.001	0.001	0.001
	0.3	0	0.007	0.007	0.007	0.001	0.001	0.001	0.000	0.000	0.000
	0.4	0	0.001	0.001	0.001	0.003	0.003	0.003	0.000	0.000	0.000
	0.5	0	0.005	0.005	0.005	0.001	0.001	0.001	0.002	0.002	0.002

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B2

Mean relative bias of loading estimates across items in group A with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	**0.227**	**0.226**	**0.226**	**0.232**	**0.232**	**0.233**	**0.234**	**0.234**	**0.234**
	0	0.2	**0.227**	**0.228**	**0.224**	**0.231**	**0.231**	**0.232**	**0.235**	**0.235**	**0.234**
	0	0.3	**0.219**	**0.220**	**0.221**	**0.229**	**0.229**	**0.229**	**0.233**	**0.233**	**0.233**
	0	0.4	**0.220**	**0.219**	**0.223**	**0.234**	**0.234**	**0.235**	**0.231**	**0.231**	**0.231**
	0	0.5	**0.222**	**0.223**	**0.221**	**0.232**	**0.233**	**0.233**	**0.236**	**0.236**	**0.236**
	0.2	0	**0.227**	**0.226**	**0.226**	**0.235**	**0.235**	**0.235**	**0.228**	**0.228**	**0.228**
	0.3	0	**0.224**	**0.223**	**0.224**	**0.236**	**0.236**	**0.236**	**0.232**	**0.232**	**0.232**
	0.4	0	**0.218**	**0.218**	**0.217**	**0.231**	**0.231**	**0.231**	**0.236**	**0.236**	**0.236**
	0.5	0	**0.225**	**0.225**	**0.226**	**0.231**	**0.231**	**0.231**	**0.231**	**0.231**	**0.231**
WLSMVPD	0	0	0.006	0.006	0.006	0.003	0.003	0.003	0.002	0.002	0.002
	0	0.2	0.005	0.005	0.005	0.001	0.001	0.001	0.002	0.002	0.002
	0	0.3	0.005	0.005	0.005	0.001	0.001	0.001	0.001	0.001	0.001
	0	0.4	0.006	0.006	0.006	0.004	0.004	0.004	0.000	0.000	0.000
	0	0.5	0.004	0.004	0.004	0.002	0.002	0.002	0.003	0.003	0.003
	0.2	0	0.008	0.008	0.008	0.005	0.005	0.005	-0.002	-0.002	-0.002
	0.3	0	0.004	0.004	0.004	0.005	0.005	0.005	0.001	0.001	0.001
	0.4	0	0.000	0.000	0.000	0.002	0.002	0.002	0.003	0.003	0.003
	0.5	0	0.007	0.007	0.007	0.003	0.003	0.003	0.000	0.000	0.000

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B3

Mean relative bias of loading estimates across complete items in group B with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	0.053	0.053	0.053	0.054	0.054	0.054	0.057	0.057	0.057
	0	0.2	0.054	0.054	0.054	0.062	0.062	0.062	0.057	0.057	0.057
	0	0.3	0.053	0.053	0.053	0.058	0.057	0.057	0.055	0.055	0.055
	0	0.4	0.055	0.055	0.055	0.054	0.054	0.054	0.056	0.056	0.056
	0	0.5	0.050	0.051	0.050	0.058	0.059	0.058	0.056	0.056	0.056
	0.2	0	0.052	0.053	0.052	0.054	0.054	0.054	0.058	0.059	0.059
	0.3	0	0.056	0.056	0.057	0.058	0.057	0.057	0.055	0.055	0.055
	0.4	0	0.055	0.055	0.055	0.060	0.060	0.060	0.060	0.060	0.060
	0.5	0	0.048	0.048	0.048	0.059	0.059	0.059	0.057	0.057	0.057
WLSMVPD	0	0	0.005	0.005	0.003	0.001	0.000	-0.001	0.002	0.001	0.000
	0	0.2	0.005	0.005	0.003	0.006	0.006	0.005	0.001	0.001	0.000
	0	0.3	0.005	0.005	0.003	0.004	0.003	0.002	0.001	0.000	-0.001
	0	0.4	0.007	0.006	0.005	0.001	0.001	0.000	0.001	0.001	0.000
	0	0.5	0.003	0.002	0.001	0.005	0.004	0.003	0.001	0.000	0.000
	0.2	0	0.003	0.003	0.003	0.001	0.001	0.001	0.003	0.003	0.003
	0.3	0	0.007	0.007	0.007	0.002	0.002	0.002	0.000	0.001	0.000
	0.4	0	0.006	0.006	0.006	0.005	0.005	0.005	0.003	0.003	0.003
	0.5	0	-0.001	0.000	0.000	0.004	0.004	0.004	0.001	0.001	0.002

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B4

Mean relative bias of loading estimates across complete items in group B with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	**0.221**	**0.221**	**0.219**	**0.233**	**0.233**	**0.232**	**0.236**	**0.235**	**0.235**
	0	0.2	**0.224**	**0.224**	**0.223**	**0.229**	**0.229**	**0.228**	**0.239**	**0.239**	**0.239**
	0	0.3	**0.219**	**0.216**	**0.216**	**0.231**	**0.231**	**0.230**	**0.231**	**0.231**	**0.231**
	0	0.4	**0.219**	**0.219**	**0.218**	**0.226**	**0.225**	**0.226**	**0.232**	**0.232**	**0.232**
	0	0.5	**0.213**	**0.212**	**0.215**	**0.231**	**0.231**	**0.231**	**0.235**	**0.234**	**0.235**
	0.2	0	**0.224**	**0.224**	**0.225**	**0.236**	**0.236**	**0.235**	**0.231**	**0.231**	**0.231**
	0.3	0	**0.222**	**0.221**	**0.220**	**0.227**	**0.227**	**0.227**	**0.233**	**0.233**	**0.233**
	0.4	0	**0.220**	**0.220**	**0.220**	**0.230**	**0.230**	**0.230**	**0.234**	**0.234**	**0.234**
	0.5	0	**0.224**	**0.224**	**0.223**	**0.238**	**0.238**	**0.239**	**0.239**	**0.239**	**0.239**
WLSMVPD	0	0	0.006	0.004	0.003	0.004	0.003	0.002	0.002	0.001	0.001
	0	0.2	0.005	0.004	0.003	0.001	0.000	-0.001	0.005	0.004	0.003
	0	0.3	0.003	0.002	0.001	0.002	0.000	0.000	0.000	-0.001	-0.001
	0	0.4	0.003	0.002	0.001	-0.002	-0.003	-0.004	0.001	0.001	0.000
	0	0.5	0.001	0.000	-0.001	0.003	0.001	0.000	0.002	0.001	0.000
	0.2	0	0.000	0.000	0.000	0.004	0.004	0.004	0.000	0.000	-0.001
	0.3	0	0.005	0.005	0.004	-0.001	-0.001	-0.001	0.001	0.001	0.001
	0.4	0	0.003	0.003	0.003	0.000	0.000	0.001	0.001	0.001	0.001
	0.5	0	0.005	0.005	0.005	0.006	0.006	0.006	0.003	0.003	0.003

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B5

Mean relative bias of loading estimates across incomplete items in group B with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	0.050	0.055	0.049	0.056	0.059	0.056	0.055	0.055	0.055
	0	0.2	0.045	0.038	0.031	0.051	0.048	0.042	0.049	0.046	0.046
	0	0.3	0.040	0.035	0.032	0.044	0.037	0.030	0.044	0.036	0.028
	0	0.4	0.026	0.017	0.006	0.030	0.022	0.011	0.029	0.023	0.013
	0	0.5	0.014	0.005	-0.011	0.020	0.013	-0.001	0.020	0.012	-0.006
	0.2	0	0.040	0.047	0.043	0.054	0.053	0.052	0.054	0.056	0.052
	0.3	0	0.073	0.074	0.092	0.060	0.065	0.072	0.046	0.045	0.049
	0.4	0	0.040	0.043	0.057	0.042	0.041	0.027	0.042	0.051	0.041
	0.5	0	0.042	0.052	0.048	0.068	0.070	0.055	0.050	0.048	0.063
WLSMVPD	0	0	0.002	0.004	-0.039	0.002	-0.001	-0.042	0.001	-0.006	-0.045
	0	0.2	0.003	-0.002	-0.036	0.003	-0.003	-0.045	0.002	-0.005	-0.043
	0	0.3	0.008	0.005	-0.030	0.004	-0.004	-0.044	0.002	-0.006	-0.049
	0	0.4	0.002	-0.004	-0.037	0.002	-0.005	-0.045	-0.001	-0.007	-0.050
	0	0.5	0.007	0.002	-0.027	0.005	0.002	-0.038	0.003	-0.003	-0.047
	0.2	0	0.000	0.001	-0.045	0.004	-0.004	-0.052	0.000	-0.008	-0.063
	0.3	0	0.031	0.026	-0.007	0.009	0.006	-0.047	-0.005	-0.015	-0.067
	0.4	0	-0.002	-0.005	-0.051	-0.005	-0.017	-0.077	-0.009	-0.011	-0.076
	0.5	0	-0.007	-0.004	-0.050	0.018	0.007	-0.058	-0.004	-0.022	-0.064

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B6

Mean relative bias of loading estimates across incomplete items in group B with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_F	DIF_T	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
rFIML	0	0	**0.223**	**0.226**	**0.204**	**0.237**	**0.242**	**0.246**	**0.235**	**0.239**	**0.253**
	0	0.2	**0.303**	**0.290**	**0.266**	**0.315**	**0.309**	**0.299**	**0.314**	**0.311**	**0.302**
	0	0.3	**0.324**	**0.302**	**0.277**	**0.333**	**0.324**	**0.298**	**0.334**	**0.324**	**0.308**
	0	0.4	**0.333**	**0.306**	**0.273**	**0.347**	**0.331**	**0.304**	**0.354**	**0.339**	**0.321**
	0	0.5	**0.340**	**0.309**	**0.268**	**0.358**	**0.336**	**0.302**	**0.366**	**0.344**	**0.313**
	0.2	0	**0.213**	**0.214**	**0.199**	**0.220**	**0.225**	**0.216**	**0.219**	**0.214**	**0.209**
	0.3	0	**0.210**	**0.206**	**0.205**	**0.195**	**0.192**	**0.176**	**0.205**	**0.197**	**0.184**
	0.4	0	**0.213**	**0.196**	**0.167**	**0.209**	**0.206**	**0.180**	**0.205**	**0.202**	**0.182**
	0.5	0	**0.164**	**0.153**	**0.129**	**0.200**	**0.174**	**0.188**	**0.202**	**0.165**	**0.151**
WLSMVPD	0	0	0.006	-0.002	-0.045	0.006	-0.006	-0.048	0.002	-0.011	-0.052
	0	0.2	0.007	0.002	-0.036	0.005	-0.006	-0.049	0.003	-0.007	-0.055
	0	0.3	0.010	0.005	-0.035	0.002	-0.009	-0.056	0.000	-0.012	-0.060
	0	0.4	0.008	-0.001	-0.046	0.000	-0.008	-0.052	0.001	-0.011	-0.057
	0	0.5	0.002	-0.002	-0.045	0.001	-0.007	-0.055	0.003	-0.008	-0.056
	0.2	0	0.014	0.011	-0.047	0.009	0.000	-0.060	0.004	-0.015	-0.068
	0.3	0	0.010	0.000	-0.044	-0.004	-0.017	-0.080	-0.001	-0.019	-0.080
	0.4	0	0.026	0.008	-0.055	0.010	-0.002	-0.068	0.003	-0.008	-0.073
	0.5	0	-0.024	-0.032	-0.091	-0.003	-0.030	-0.060	0.010	-0.028	-0.088

Note: rFIML: robust FIML, WLSMVPDL: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings, complete: complete data condition, 30% miss: 30% of missing data are imposed on the incomplete items in group B, 50% miss: 50% of missing data are imposed on the incomplete items in group B.

Table B7

Mean relative bias of loadings' standard errors across items in group A with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_L	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.056	0.056	0.055	0.071	0.071	0.071	0.083	0.083	0.083
	0.2	0	0.065	0.065	0.067	0.060	0.060	0.060	0.060	0.060	0.060
	0.3	0	0.059	0.059	0.059	0.058	0.058	0.059	0.073	0.073	0.073
	0.4	0	0.067	0.067	0.066	0.068	0.068	0.068	0.055	0.055	0.055
	0.5	0	0.061	0.061	0.063	0.055	0.055	0.055	0.051	0.051	0.051
	0	0.2	0.039	0.039	0.039	0.059	0.059	0.059	0.070	0.070	0.070
	0	0.3	0.074	0.074	0.074	0.030	0.030	0.030	0.041	0.041	0.041
	0	0.4	0.051	0.051	0.051	0.070	0.070	0.070	0.067	0.067	0.067
	0	0.5	0.052	0.052	0.052	0.058	0.058	0.058	0.043	0.043	0.043
rFIML	0	0	0.002	0.002	0.002	0.012	0.012	0.012	0.021	0.021	0.021
	0.2	0	0.012	0.012	0.013	0.003	0.003	0.003	0.000	0.000	0.000
	0.3	0	0.004	0.004	0.004	-0.001	-0.001	-0.001	0.011	0.011	0.011
	0.4	0	0.015	0.015	0.014	0.01	0.010	0.010	-0.004	-0.004	-0.004
	0.5	0	0.011	0.011	0.013	-0.003	-0.003	-0.003	-0.009	-0.009	-0.009
	0	0.2	-0.011	-0.011	-0.011	0.000	0.000	0.000	0.010	0.010	0.01
	0	0.3	0.021	0.021	0.021	-0.025	-0.026	-0.026	-0.018	-0.018	-0.018
	0	0.4	0.000	0.000	0.000	0.010	0.010	0.010	0.007	0.007	0.007
	0	0.5	0.000	0.000	0.000	0.000	0.000	0.000	-0.017	-0.017	-0.017
WLSMVPD	0	0	-0.056	-0.056	-0.056	-0.021	-0.021	-0.021	0.001	0.001	0.001
	0.2	0	-0.062	-0.062	-0.062	-0.033	-0.033	-0.033	-0.023	-0.023	-0.023
	0.3	0	-0.068	-0.068	-0.068	-0.028	-0.028	-0.029	-0.020	-0.020	-0.020
	0.4	0	-0.055	-0.054	-0.055	-0.020	-0.020	-0.020	-0.018	-0.018	-0.018
	0.5	0	-0.056	-0.056	-0.057	-0.042	-0.042	-0.041	-0.027	-0.027	-0.027
	0	0.2	-0.067	-0.067	-0.067	-0.025	-0.025	-0.025	-0.016	-0.016	-0.016
	0	0.3	-0.057	-0.057	-0.057	-0.047	-0.047	-0.047	-0.038	-0.038	-0.038
	0	0.4	-0.069	-0.069	-0.069	-0.037	-0.037	-0.037	-0.015	-0.015	-0.015
	0	0.5	-0.068	-0.068	-0.068	-0.029	-0.029	-0.029	-0.037	-0.037	-0.037

Note: rFIML: robust FIML, WLSMVPDL: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

Table B8

Mean relative bias of loadings' standard errors across items in group A with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_L	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.047	0.048	0.040	0.017	0.017	0.017	0.011	0.010	0.011
	0.2	0	0.024	0.030	0.027	0.006	0.005	0.004	0.008	0.008	0.009
	0.3	0	0.027	0.026	0.033	0.005	0.003	0.003	0.010	0.010	0.009
	0.4	0	0.038	0.044	0.047	0.022	0.023	0.023	0.023	0.022	0.023
	0.5	0	0.028	0.024	0.024	0.011	0.010	0.012	0.036	0.036	0.035
	0	0.2	0.036	0.035	0.036	0.022	0.022	0.023	0.019	0.019	0.019
	0	0.3	0.055	0.053	0.050	0.013	0.013	0.013	0.017	0.017	0.017
	0	0.4	0.016	0.017	0.016	0.019	0.019	0.019	0.018	0.018	0.018
	0	0.5	0.033	0.030	0.033	0.018	0.018	0.018	0.022	0.022	0.022
rFIML	0	0	0.045	0.047	0.038	0.003	0.004	0.003	-0.007	-0.008	-0.007
	0.2	0	0.024	0.030	0.027	-0.005	-0.006	-0.008	-0.009	-0.009	-0.008
	0.3	0	0.025	0.023	0.028	-0.008	-0.009	-0.010	-0.007	-0.007	-0.008
	0.4	0	0.039	0.046	0.048	0.010	0.011	0.011	0.007	0.006	0.006
	0.5	0	0.027	0.023	0.022	-0.002	-0.002	-0.001	0.016	0.016	0.016
	0	0.2	0.036	0.036	0.036	0.010	0.010	0.011	0.004	0.004	0.004
	0	0.3	0.056	0.053	0.05-	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	0	0.4	0.016	0.017	0.017	0.007	0.007	0.007	0.001	0.001	0.001
	0	0.5	0.031	0.029	0.032	0.006	0.006	0.005	0.006	0.006	0.006
WLSMVPD	0	0	-0.042	-0.042	-0.042	-0.040	-0.040	-0.040	-0.025	-0.025	-0.025
	0.2	0	-0.067	-0.067	-0.067	-0.043	-0.043	-0.043	-0.026	-0.026	-0.026
	0.3	0	-0.056	-0.056	-0.056	-0.036	-0.036	-0.036	-0.029	-0.029	-0.029
	0.4	0	-0.058	-0.058	-0.058	-0.023	-0.023	-0.023	-0.009	-0.009	-0.009
	0.5	0	-0.066	-0.066	-0.066	-0.039	-0.039	-0.039	0.001	0.001	0.001
	0	0.2	-0.058	-0.058	-0.058	-0.024	-0.024	-0.024	-0.012	-0.012	-0.013
	0	0.3	-0.050	-0.050	-0.050	-0.044	-0.044	-0.044	-0.015	-0.015	-0.015
	0	0.4	-0.066	-0.066	-0.066	-0.030	-0.030	-0.030	-0.008	-0.008	-0.008
	0	0.5	-0.061	-0.061	-0.062	-0.026	-0.026	-0.026	-0.018	-0.017	-0.017

Note: rFIML: robust FIML, WLSMVPDL: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

Table B9

Mean relative bias of loadings' standard errors across complete items in group B with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_I	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.054	0.056	0.052	0.050	0.046	0.050	0.069	0.068	0.061
	0.2	0	0.056	0.051	0.047	0.072	0.067	0.066	0.078	0.076	0.078
	0.3	0	0.027	0.025	0.023	0.059	0.049	0.045	0.063	0.058	0.063
	0.4	0	0.059	0.052	0.051	0.069	0.062	0.061	0.049	0.045	0.041
	0.5	0	0.066	0.063	0.062	0.065	0.059	0.060	0.049	0.045	0.041
	0	0.2	0.038	0.036	0.031	0.040	0.038	0.039	0.070	0.070	0.068
	0	0.3	0.046	0.047	0.044	0.046	0.047	0.045	0.056	0.055	0.056
	0	0.4	0.054	0.055	0.052	0.024	0.024	0.024	0.063	0.062	0.062
	0	0.5	0.036	0.036	0.035	0.042	0.043	0.042	0.056	0.055	0.057
rFIML	0	0	0.003	0.009	0.008	-0.007	-0.008	-0.003	0.007	0.009	0.004
	0.2	0	0.004	0.004	0.002	0.012	0.010	0.011	0.016	0.016	0.020
	0.3	0	-0.024	-0.022	-0.021	0.001	-0.005	-0.007	0.003	0.001	0.007
	0.4	0	0.008	0.005	0.008	0.011	0.008	0.009	-0.01-	-0.012	-0.012
	0.5	0	0.015	0.018	0.020	0.006	0.004	0.007	-0.012	-0.012	-0.014
	0	0.2	-0.005	-0.003	-0.007	-0.012	-0.012	-0.010	0.013	0.014	0.013
	0	0.3	0.006	0.008	0.007	-0.004	-0.002	-0.003	0.003	0.002	0.004
	0	0.4	0.017	0.019	0.016	-0.023	-0.023	-0.023	0.009	0.009	0.010
	0	0.5	0.002	0.003	0.002	-0.006	-0.004	-0.005	0.005	0.004	0.006
WLSMVPD	0	0	-0.066	-0.071	-0.080	-0.034	-0.039	-0.037	-0.017	-0.021	-0.028
	0.2	0	-0.063	-0.073	-0.085	-0.025	-0.032	-0.036	0.005	0.005	-0.001
	0.3	0	-0.093	-0.097	-0.105	-0.027	-0.037	-0.037	-0.016	-0.023	-0.016
	0.4	0	-0.060	-0.073	-0.072	-0.022	-0.028	-0.031	-0.034	-0.039	-0.041
	0.5	0	-0.052	-0.057	-0.064	-0.030	-0.038	-0.037	-0.031	-0.033	-0.036
	0	0.2	-0.076	-0.078	-0.084	-0.035	-0.037	-0.036	-0.014	-0.015	-0.017
	0	0.3	-0.063	-0.064	-0.069	-0.043	-0.042	-0.045	-0.021	-0.023	-0.021
	0	0.4	-0.054	-0.053	-0.057	-0.057	-0.057	-0.059	-0.014	-0.016	-0.016
	0	0.5	-0.066	-0.067	-0.067	-0.039	-0.037	-0.039	-0.024	-0.025	-0.024

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

Table B10

Mean relative bias of loadings' standard errors across complete items in group B with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_L	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.035	0.038	0.023	0.003	0.002	0.006	0.008	0.007	0.008
	0.2	0	0.048	0.037	0.027	0.027	0.021	0.019	0.032	0.032	0.030
	0.3	0	0.037	0.035	0.025	0.044	0.027	0.031	0.033	0.035	0.029
	0.4	0	0.026	0.022	0.028	0.005	0.002	-0.004	0.025	0.017	0.017
	0.5	0	0.024	0.016	0.020	0.004	-0.004	-0.003	0.023	0.021	0.018
	0	0.2	0.032	0.029	0.022	-0.013	-0.015	-0.013	0.016	0.014	0.011
	0	0.3	0.047	0.042	0.041	0.014	0.013	0.009	0.019	0.018	0.018
	0	0.4	0.024	0.021	0.023	0.004	0.002	0.004	-0.004	-0.006	-0.007
	0	0.5	0.016	0.017	0.013	0.022	0.022	0.020	0.022	0.023	0.022
rFIML	0	0	0.035	0.042	0.031	-0.01	-0.009	-0.001	-0.009	-0.008	-0.005
	0.2	0	0.044	0.040	0.036	0.011	0.009	0.011	0.009	0.013	0.013
	0.3	0	0.030	0.036	0.031	0.022	0.011	0.020	0.009	0.015	0.013
	0.4	0	0.016	0.021	0.034	-0.015	-0.012	-0.013	0.000	-0.002	0.002
	0.5	0	0.011	0.014	0.025	-0.020	-0.020	-0.014	-0.006	-0.001	0.000
	0	0.2	0.038	0.039	0.035	-0.022	-0.022	-0.018	0.003	0.002	0.002
	0	0.3	0.059	0.058	0.058	0.010	0.010	0.008	0.009	0.011	0.012
	0	0.4	0.045	0.044	0.048	0.004	0.002	0.005	-0.012	-0.012	-0.013
	0	0.5	0.041	0.042	0.039	0.024	0.023	0.022	0.016	0.017	0.017
WLSMVPD	0	0	-0.058	-0.061	-0.078	-0.044	-0.043	-0.045	-0.019	-0.020	-0.021
	0.2	0	-0.056	-0.059	-0.059	-0.026	-0.032	-0.028	-0.007	-0.005	-0.008
	0.3	0	-0.048	-0.057	-0.060	-0.014	-0.024	-0.023	-0.016	-0.012	-0.015
	0.4	0	-0.060	-0.057	-0.058	-0.039	-0.041	-0.047	-0.023	-0.023	-0.023
	0.5	0	-0.065	-0.070	-0.073	-0.045	-0.042	-0.042	-0.023	-0.020	-0.018
	0	0.2	-0.065	-0.069	-0.072	-0.061	-0.060	-0.061	-0.009	-0.010	-0.011
	0	0.3	-0.051	-0.052	-0.052	-0.025	-0.025	-0.030	-0.016	-0.012	-0.011
	0	0.4	-0.061	-0.066	-0.068	-0.038	-0.041	-0.041	-0.040	-0.039	-0.040
	0	0.5	-0.057	-0.058	-0.057	-0.019	-0.019	-0.019	0.002	0.003	0.002

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

Table B11

Mean relative bias of loadings' standard errors across incomplete items in group B with symmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_L	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.082	0.043	0.044	0.054	0.051	0.058	0.052	0.048	0.011
	0.2	0	0.059	0.068	0.023	0.023	0.034	0.015	0.083	0.087	0.050
	0.3	0	0.057	0.043	0.048	0.046	0.078	0.033	0.069	0.079	0.037
	0.4	0	0.069	0.044	0.024	0.033	0.008	0.007	0.064	0.064	0.037
	0.5	0	0.080	0.088	0.056	0.042	0.044	0.051	0.049	0.053	0.057
	0	0.2	0.026	0.004	0.029	0.041	0.026	0.013	0.045	0.052	0.008
	0	0.3	0.016	0.033	-0.011	-0.005	-0.005	0.005	0.024	0.010	0.023
	0	0.4	-0.001	-0.011	-0.035	0.014	0.002	0.023	0.015	-0.005	0.005
	0	0.5	-0.031	-0.037	-0.018	-0.003	-0.013	-0.027	-0.011	-0.006	-0.016
rFIML	0	0	0.030	-0.003	0.006	-0.004	-0.003	0.010	-0.007	-0.006	-0.036
	0.2	0	0.007	0.019	-0.018	-0.033	-0.020	-0.030	0.021	0.029	-0.004
	0.3	0	0.003	-0.005	0.003	-0.013	0.023	-0.015	0.008	0.021	-0.014
	0.4	0	0.016	-0.004	-0.013	-0.024	-0.042	-0.039	0.005	0.006	-0.012
	0.5	0	0.027	0.040	0.017	-0.015	-0.011	0.002	-0.010	-0.002	0.005
	0	0.2	0.017	-0.004	0.026	0.020	0.007	-0.002	0.021	0.028	-0.013
	0	0.3	0.018	0.034	-0.009	-0.011	-0.010	0.000	0.014	0.001	0.015
	0	0.4	0.011	-0.001	-0.024	0.017	0.005	0.027	0.014	-0.006	0.004
	0	0.5	-0.014	-0.019	0.001	0.003	-0.009	-0.022	-0.007	-0.003	-0.014
WLSMVPD	0	0	-0.043	-0.089	**-0.134**	-0.037	-0.075	-0.058	-0.022	-0.033	-0.064
	0.2	0	-0.049	-0.063	**-0.136**	-0.055	-0.046	-0.065	-0.013	-0.004	-0.021
	0.3	0	-0.066	-0.094	**-0.124**	-0.048	-0.026	-0.086	-0.019	-0.021	-0.039
	0.4	0	-0.029	-0.089	**-0.125**	-0.055	-0.081	-0.090	-0.022	-0.035	-0.049
	0.5	0	-0.052	-0.078	**-0.121**	-0.055	-0.064	-0.058	-0.028	-0.037	-0.046
	0	0.2	-0.072	**-0.108**	**-0.110**	-0.020	-0.041	-0.056	-0.008	0.005	-0.043
	0	0.3	-0.062	-0.070	**-0.132**	-0.047	-0.055	-0.059	-0.020	-0.034	-0.038
	0	0.4	-0.074	**-0.105**	**-0.138**	-0.022	-0.041	-0.034	-0.009	-0.029	-0.024
	0	0.5	-0.094	**-0.113**	**-0.129**	-0.039	-0.052	-0.082	-0.029	-0.026	-0.047

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

Table B12

Mean relative bias of loadings' standard errors across incomplete items in group B with asymmetric distributed thresholds

			N = 300			N = 600			N = 1000		
Estimator	DIF_T	DIF_L	complete	30% miss	50% miss	complete	30% miss	50% miss	complete	30% miss	50% miss
FIML	0	0	0.023	0.059	0.076	0.043	0.052	0.090	0.009	0.018	0.020
	0.2	0	**0.115**	**0.185**	**0.185**	**0.105**	**0.121**	**0.173**	0.039	0.040	0.06-
	0.3	0	**0.169**	**0.188**	**0.267**	**0.106**	**0.107**	**0.148**	**0.102**	**0.106**	**0.107**
	0.4	0	**0.189**	**0.182**	**0.229**	**0.171**	**0.184**	**0.162**	**0.164**	**0.172**	**0.187**
	0.5	0	**0.222**	**0.239**	**0.223**	**0.141**	**0.156**	**0.134**	**0.117**	**0.122**	**0.105**
	0	0.2	-0.030	-0.009	-0.037	0.036	0.037	0.047	0.014	0.014	0.007
	0	0.3	-0.045	-0.043	-0.021	-0.038	-0.030	-0.042	0.012	0.012	0.002
	0	0.4	-0.017	0.004	-0.035	-0.041	-0.046	-0.026	-0.032	-0.025	-0.020
	0	0.5	-0.041	-0.027	-0.052	-0.024	-0.017	-0.034	0.015	0.003	-0.004
rFIML	0	0	0.025	0.049	0.052	0.03	0.020	0.041	-0.009	-0.018	-0.034
	0.2	0	0.056	**0.118**	**0.118**	0.036	0.042	0.085	-0.032	-0.041	-0.030
	0.3	0	0.086	**0.105**	**0.177**	0.013	0.009	0.045	0.005	0.002	0.000
	0.4	0	0.084	0.081	**0.134**	0.054	0.063	0.048	0.043	0.047	0.062
	0.5	0	0.099	**0.121**	**0.122**	0.010	0.026	0.018	-0.015	-0.010	-0.018
	0	0.2	0.007	0.022	-0.014	0.060	0.051	0.050	0.032	0.023	0.007
	0	0.3	-0.001	-0.006	0.012	-0.010	-0.007	-0.026	0.035	0.029	0.011
	0	0.4	0.028	0.045	0.001	-0.012	-0.022	-0.005	-0.008	-0.005	-0.005
	0	0.5	-0.006	0.006	-0.023	0.001	0.006	-0.012	0.033	0.018	0.008
WLSMVPD	0	0	-0.052	-0.067	**-0.145**	-0.002	-0.015	-0.035	-0.030	-0.047	-0.067
	0.2	0	-0.038	-0.038	**-0.111**	-0.005	-0.016	-0.034	-0.047	-0.066	-0.062
	0.3	0	-0.044	-0.083	-0.087	-0.038	-0.051	-0.057	-0.017	-0.021	-0.034
	0.4	0	-0.048	-0.075	-0.072	-0.001	0.011	-0.047	0.018	0.001	-0.014
	0.5	0	-0.043	-0.066	-0.083	-0.021	-0.037	-0.082	-0.035	-0.037	-0.043
	0	0.2	-0.074	-0.082	**-0.142**	0.011	-0.019	-0.025	0.009	0.006	-0.026
	0	0.3	-0.077	-0.086	**-0.116**	-0.052	-0.052	-0.078	0.006	-0.005	-0.018
	0	0.4	-0.059	-0.060	**-0.107**	-0.060	-0.071	-0.058	-0.030	-0.029	-0.040
	0	0.5	-0.080	-0.085	**-0.139**	-0.040	-0.046	-0.076	0.010	-0.005	-0.018

Note: rFIML: robust FIML, WLSMVPD: mean and variance adjusted weight least squared with pairwise deletion method, DIF_T: amount of non-invariance in thresholds, DIF_L: amount of non-invariance in loadings.

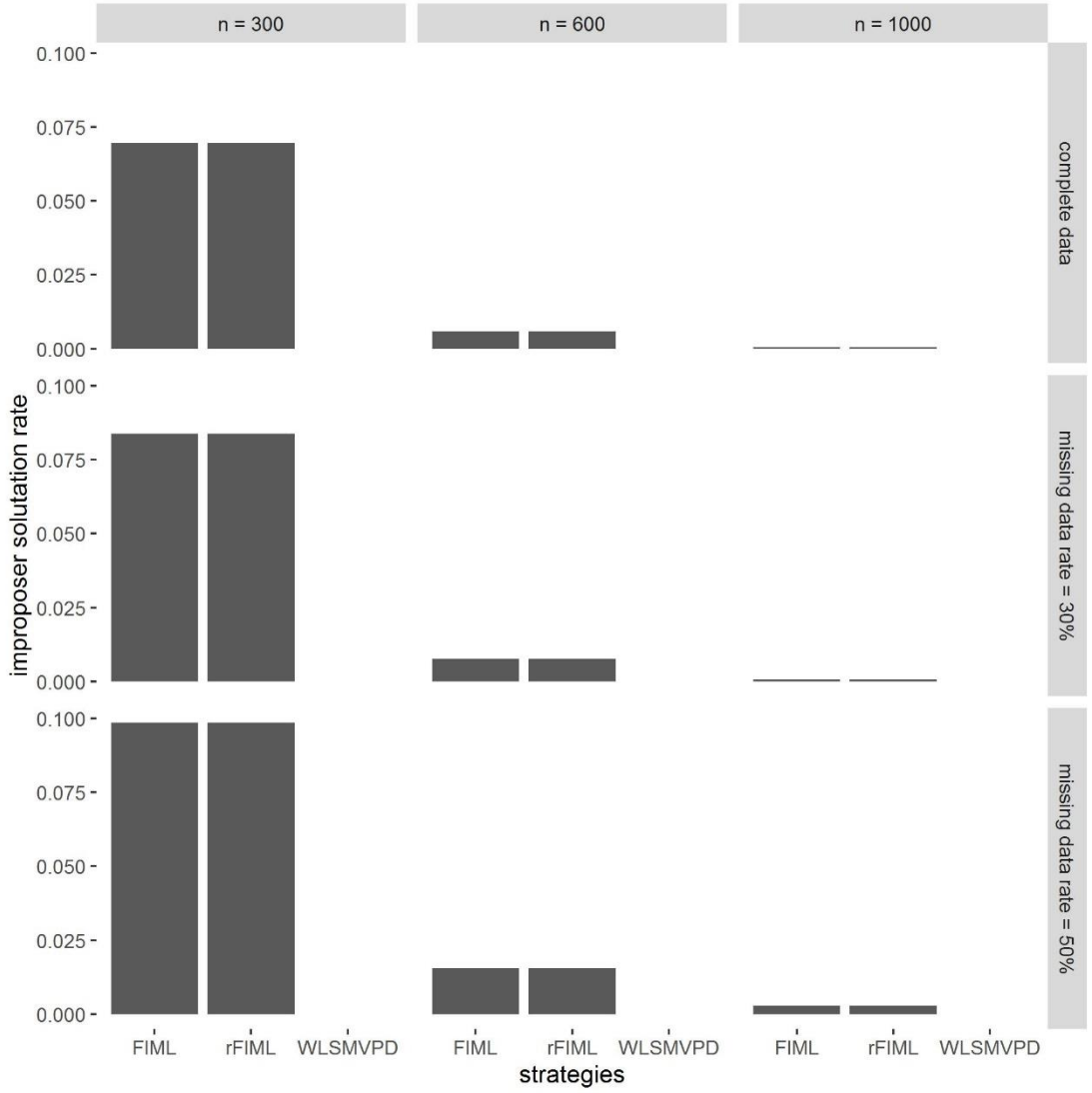

Figure B1

Improper solution rates of having standardize loadings larger than 1

Note: FIML: continuous full information likelihood methods, rFIML: robust continuous full information likelihood methods, WLSMVPD: *weighted least squares* means and variance adjusted estimators plus pairwise deletion. Given FIML & rFIML have the same point estimates, the improper solution rates of them are always equal.

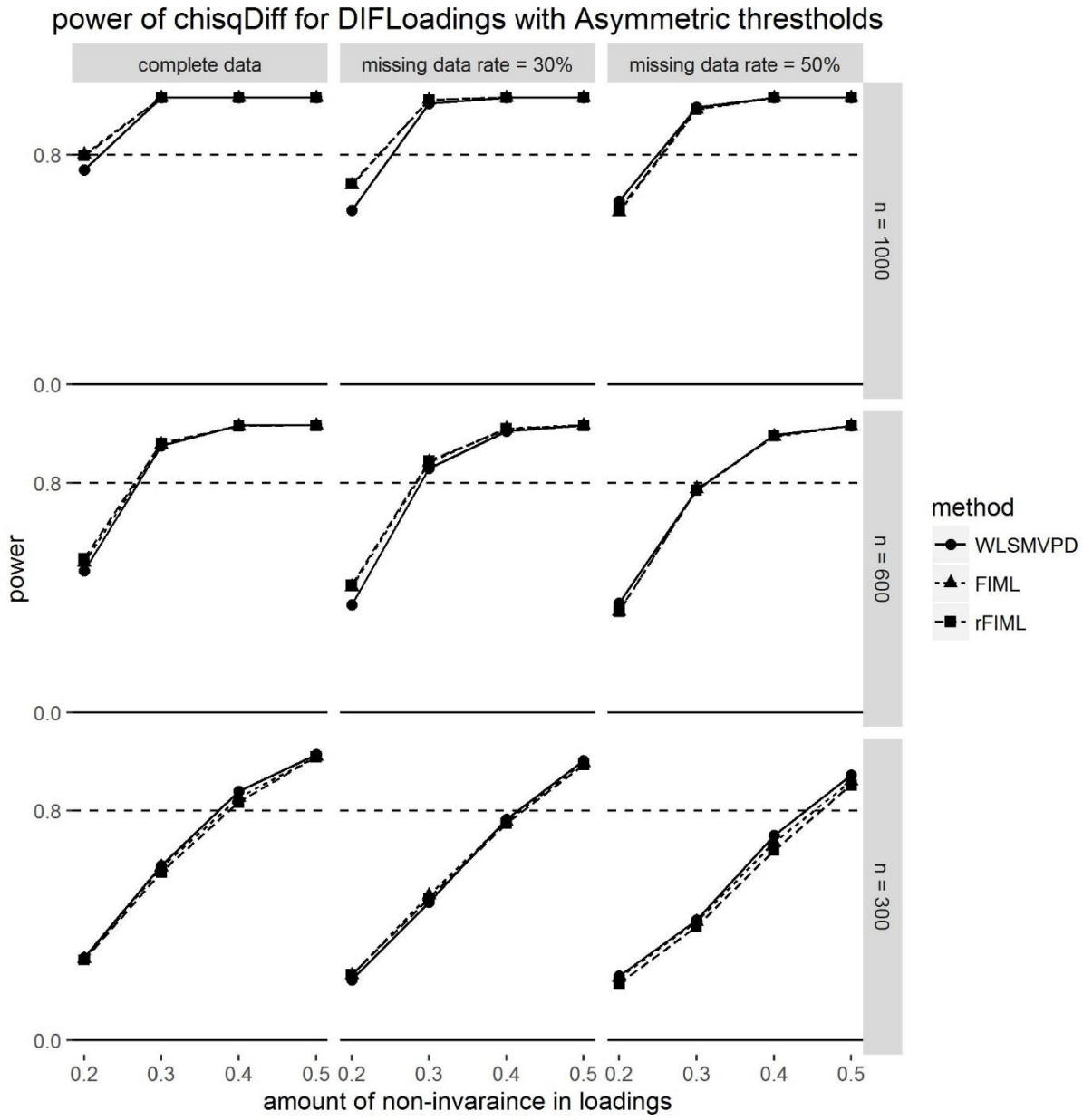

Figure B2

Power of the $\Delta\chi^2$ tests on defecting non-invariant loadings when thresholds are asymmetric distributed.

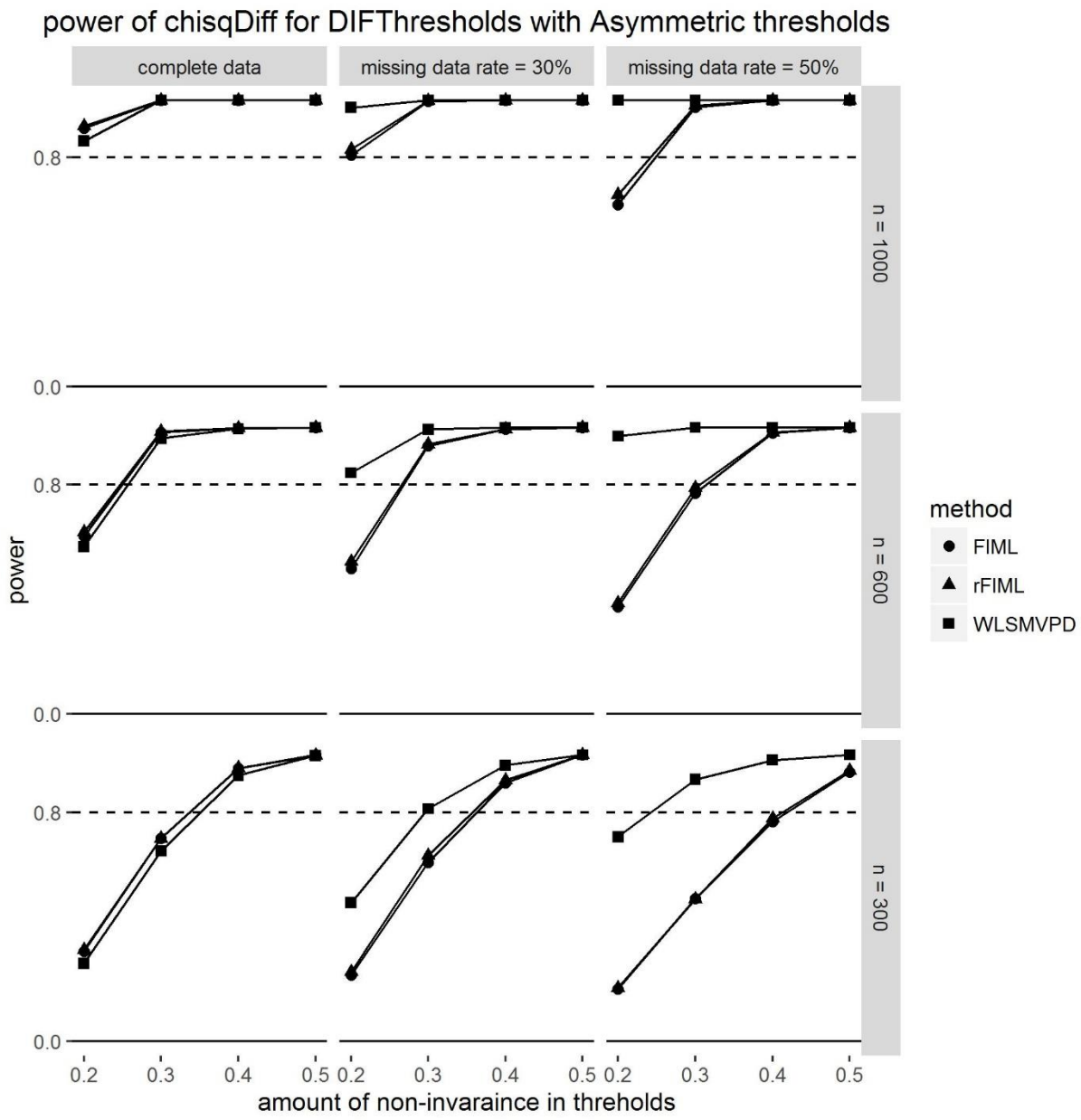

Figure B3

Power of the $\Delta\chi^2$ tests on defecting non-invariant thresholds when thresholds are asymmetric distributed.

Appendix C

Table C1

Perfect recovery rate of methods in loading non-invariant conditions (α = 0.05/cutoff of the modification indices is set as 3.841).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.266	0.270	**0.384**	0.220
	30%		0.196	0.198	**0.288**	0.240
	50%		0.122	0.126	0.204	**0.256**
1000	Complete		0.682	0.678	**0.744**	0.480
	30%		0.548	0.548	**0.628**	0.496
	50%		0.340	0.362	0.492	**0.576**
2000	Complete		0.872	0.874	**0.884**	0.681
	30%		0.838	0.842	**0.876**	0.690
	50%		0.764	0.766	**0.838**	0.724
400	Complete	large	**0.874**	0.870	0.868	0.670
	30%		0.810	0.810	**0.840**	0.607
	50%		0.690	0.692	**0.738**	0.615
1000	Complete		**0.902**	**0.902**	0.886	0.728
	30%		**0.892**	0.890	0.886	0.726
	50%		**0.882**	0.880	0.880	0.729
2000	Complete		**0.900**	0.898	0.868	0.698
	30%		**0.902**	**0.902**	0.862	0.714
	50%		**0.896**	0.890	0.878	0.720
400	Complete	mixed	**0.816**	0.816	0.812	0.613
	30%		0.716	0.716	**0.730**	0.563
	50%		0.608	0.608	**0.646**	0.557
1000	Complete		**0.908**	0.906	0.886	0.743
	30%		**0.880**	0.874	**0.876**	0.747
	50%		**0.848**	0.844	**0.848**	0.756
2000	Complete		**0.906**	0.902	0.858	0.737
	30%		**0.896**	0.898	0.866	0.738
	50%		**0.904**	0.902	0.874	0.742
400	Complete	nonuniform	**0.892**	0.890	0.852	0.497
	30%		**0.846**	0.846	0.824	0.288
	50%		0.746	0.750	**0.776**	0.078
1000	Complete		0.876	0.870	0.868	0.616
	30%		0.882	0.880	0.864	0.476
	50%		**0.870**	0.868	0.856	0.092
2000	Complete		**0.904**	0.904	0.874	0.672
	30%		**0.904**	0.902	0.880	0.628
	50%		**0.898**	0.896	0.892	0.142

Table C2

Model level type I error rate in loading non-invariant conditions (α = 0.05/cutoff of the modification indices is set as 3.841).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.148	0.154	0.132	0.308
	30%		0.140	0.148	0.134	0.284
	50%		0.122	0.126	0.128	0.278
1000	Complete		0.114	0.118	0.118	0.300
	30%		0.122	0.124	0.106	0.290
	50%		0.130	0.130	0.104	0.282
2000	Complete		0.110	0.108	0.108	0.285
	30%		0.114	0.110	0.108	0.282
	50%		0.118	0.118	0.100	0.258
400	Complete	large	0.100	0.102	0.120	0.290
	30%		0.104	0.110	0.112	0.298
	50%		0.102	0.104	0.116	0.278
1000	Complete		0.098	0.098	0.114	0.272
	30%		0.108	0.110	0.112	0.274
	50%		0.106	0.108	0.116	0.265
2000	Complete		0.100	0.102	0.132	0.302
	30%		0.098	0.098	0.138	0.286
	50%		0.104	0.110	0.122	0.280
400	Complete	mixed	0.098	0.100	0.132	0.300
	30%		0.104	0.108	0.144	0.306
	50%		0.106	0.110	0.116	0.286
1000	Complete		0.090	0.092	0.114	0.248
	30%		0.096	0.102	0.110	0.236
	50%		0.098	0.104	0.116	0.226
2000	Complete		0.094	0.098	0.142	0.263
	30%		0.104	0.102	0.134	0.262
	50%		0.096	0.098	0.126	0.258
400	Complete	nonuniform	0.084	0.086	0.106	0.364
	30%		0.098	0.100	0.132	0.324
	50%		0.114	0.116	0.134	0.296
1000	Complete		0.124	0.130	0.132	0.366
	30%		0.118	0.120	0.134	0.348
	50%		0.126	0.128	0.142	0.304
2000	Complete		0.096	0.096	0.126	0.328
	30%		0.096	0.098	0.120	0.330
	50%		0.102	0.104	0.108	0.294

Table C3

Model level type II error rate in loading non-invariant conditions (α = 0.05/cutoff of the modification indices is set as 3.841).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.692	0.688	0.548	0.714
	30%		0.772	0.772	0.656	0.718
	50%		0.872	0.868	0.744	0.690
1000	Complete		0.260	0.260	0.172	0.364
	30%		0.410	0.410	0.304	0.362
	50%		0.642	0.624	0.446	0.266
2000	Complete		0.020	0.020	0.008	0.076
	30%		0.066	0.066	0.020	0.058
	50%		0.152	0.150	0.074	0.030
400	Complete	large	0.038	0.040	0.024	0.091
	30%		0.104	0.100	0.062	0.185
	50%		0.248	0.244	0.172	0.209
1000	Complete		0.000	0.000	0.000	0.000
	30%		0.002	0.002	0.002	0.000
	50%		0.014	0.014	0.004	0.014
2000	Complete		0.000	0.000	0.000	0.000
	30%		0.000	0.000	0.000	0.000
	50%		0.000	0.000	0.000	0.000
400	Complete	mixed	0.092	0.090	0.074	0.152
	30%		0.206	0.204	0.156	0.218
	50%		0.334	0.332	0.268	0.251
1000	Complete		0.004	0.004	0.004	0.016
	30%		0.032	0.030	0.018	0.024
	50%		0.066	0.064	0.042	0.036
2000	Complete		0.000	0.000	0.000	0.000
	30%		0.000	0.000	0.000	0.000
	50%		0.000	0.000	0.000	0.000
400	Complete	nonuniform	0.038	0.038	0.056	0.240
	30%		0.078	0.078	0.064	0.584
	50%		0.198	0.192	0.112	0.904
1000	Complete		0.000	0.000	0.000	0.044
	30%		0.000	0.000	0.002	0.302
	50%		0.008	0.008	0.002	0.870
2000	Complete		0.000	0.000	0.000	0.000
	30%		0.000	0.000	0.000	0.068
	50%		0.000	0.000	0.000	0.814

Table C4

Item level type I error rate in loading non-invariant conditions (α = 0.05/cutoff of the modification indices is set as 3.841).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.058	0.061	0.054	0.129
	30%		0.053	0.057	0.056	0.118
	50%		0.044	0.046	0.051	0.117
1000	Complete		0.046	0.047	0.049	0.136
	30%		0.049	0.049	0.042	0.130
	50%		0.053	0.052	0.045	0.131
2000	Complete		0.039	0.039	0.041	0.118
	30%		0.041	0.040	0.039	0.120
	50%		0.042	0.041	0.039	0.103
400	Complete	large	0.033	0.034	0.047	0.131
	30%		0.035	0.037	0.044	0.143
	50%		0.035	0.035	0.046	0.126
1000	Complete		0.033	0.033	0.043	0.102
	30%		0.037	0.037	0.043	0.109
	50%		0.037	0.037	0.042	0.106
2000	Complete		0.037	0.038	0.057	0.111
	30%		0.035	0.035	0.057	0.108
	50%		0.039	0.041	0.051	0.103
400	Complete	mixed	0.035	0.035	0.049	0.118
	30%		0.037	0.039	0.054	0.131
	50%		0.040	0.041	0.045	0.115
1000	Complete		0.031	0.032	0.044	0.094
	30%		0.035	0.037	0.043	0.090
	50%		0.034	0.036	0.046	0.089
2000	Complete		0.034	0.035	0.057	0.096
	30%		0.037	0.037	0.051	0.095
	50%		0.035	0.035	0.050	0.096
400	Complete	nonuniform	0.028	0.029	0.045	0.135
	30%		0.033	0.034	0.051	0.119
	50%		0.039	0.039	0.051	0.109
1000	Complete		0.043	0.045	0.053	0.145
	30%		0.041	0.041	0.053	0.138
	50%		0.045	0.046	0.057	0.115
2000	Complete		0.033	0.033	0.049	0.129
	30%		0.033	0.034	0.045	0.127
	50%		0.035	0.035	0.041	0.112

Table C5

Item level power in loading non-invariant conditions (α = 0.05/cutoff of the modification indices is set as 3.841).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.494	0.501	0.631	0.451
	30%		0.424	0.425	0.538	0.445
	50%		0.312	0.321	0.435	0.470
1000	Complete		0.832	0.833	0.895	0.746
	30%		0.732	0.732	0.824	0.741
	50%		0.555	0.568	0.722	0.820
2000	Complete		0.989	0.989	0.996	0.949
	30%		0.964	0.964	0.990	0.962
	50%		0.909	0.910	0.960	0.982
400	Complete	large	0.978	0.977	0.987	0.936
	30%		0.943	0.944	0.966	0.874
	50%		0.862	0.864	0.907	0.857
1000	Complete		1.000	1.000	1.000	1.000
	30%		0.999	0.999	0.999	1.000
	50%		0.993	0.993	0.998	0.989
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	1.000
	50%		1.000	1.000	1.000	1.000
400	Complete	mixed	0.954	0.955	0.963	0.920
	30%		0.895	0.897	0.922	0.873
	50%		0.824	0.824	0.861	0.860
1000	Complete		0.998	0.998	0.998	0.992
	30%		0.984	0.985	0.991	0.988
	50%		0.967	0.968	0.979	0.981
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	1.000
	50%		1.000	1.000	1.000	1.000
400	Complete	nonuniform	0.981	0.981	0.972	0.880
	30%		0.961	0.961	0.968	0.701
	50%		0.900	0.903	0.944	0.512
1000	Complete		1.000	1.000	1.000	0.978
	30%		1.000	1.000	0.999	0.848
	50%		0.996	0.996	0.999	0.565
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	0.966
	50%		1.000	1.000	1.000	0.593

Table C6

Item level type I error rate in loading non-invariant conditions (α = 0.01/cutoff of the modification indices is set as 6.635).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.019	0.019	0.012	0.055
	30%		0.019	0.019	0.009	0.053
	50%		0.015	0.015	0.011	0.058
1000	Complete		0.020	0.021	0.011	0.067
	30%		0.016	0.016	0.009	0.072
	50%		0.018	0.018	0.009	0.067
2000	Complete		0.011	0.011	0.009	0.054
	30%		0.009	0.010	0.009	0.057
	50%		0.009	0.009	0.008	0.043
400	Complete	large	0.005	0.006	0.011	0.078
	30%		0.009	0.008	0.009	0.080
	50%		0.012	0.012	0.008	0.076
1000	Complete		0.006	0.006	0.009	0.037
	30%		0.008	0.008	0.008	0.039
	50%		0.007	0.007	0.009	0.045
2000	Complete		0.011	0.011	0.011	0.041
	30%		0.010	0.011	0.009	0.036
	50%		0.011	0.011	0.013	0.039
400	Complete	mixed	0.009	0.009	0.009	0.054
	30%		0.009	0.010	0.007	0.067
	50%		0.009	0.009	0.007	0.056
1000	Complete		0.006	0.007	0.008	0.027
	30%		0.007	0.008	0.009	0.027
	50%		0.006	0.007	0.009	0.032
2000	Complete		0.003	0.003	0.012	0.032
	30%		0.003	0.003	0.011	0.031
	50%		0.003	0.003	0.011	0.029
400	Complete	nonuniform	0.008	0.008	0.011	0.060
	30%		0.010	0.010	0.011	0.049
	50%		0.009	0.011	0.009	0.031
1000	Complete		0.010	0.010	0.014	0.059
	30%		0.008	0.008	0.015	0.056
	50%		0.011	0.009	0.015	0.046
2000	Complete		0.006	0.006	0.009	0.049
	30%		0.005	0.006	0.009	0.050
	50%		0.006	0.006	0.006	0.040

Table C7

Item level power in loading non-invariant conditions (α = 0.01/cutoff of the modification indices is set as 6.635).

Sample size	Missing rate	Type of DIF	FIML	rFIML	WLSMVMI	WLSMVPD
400	Complete	small	0.235	0.248	0.373	0.247
	30%		0.179	0.187	0.284	0.250
	50%		0.120	0.127	0.222	0.227
1000	Complete		0.601	0.613	0.744	0.595
	30%		0.472	0.484	0.608	0.564
	50%		0.307	0.316	0.476	0.627
2000	Complete		0.948	0.948	0.978	0.907
	30%		0.888	0.890	0.945	0.907
	50%		0.745	0.744	0.859	0.949
400	Complete	large	0.928	0.928	0.960	0.881
	30%		0.840	0.848	0.887	0.762
	50%		0.665	0.676	0.773	0.663
1000	Complete		0.999	0.999	1.000	0.998
	30%		0.998	0.998	0.996	0.994
	50%		0.973	0.973	0.984	0.972
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	1.000
	50%		1.000	1.000	1.000	1.000
400	Complete	mixed	0.861	0.866	0.884	0.854
	30%		0.773	0.775	0.827	0.751
	50%		0.649	0.660	0.736	0.708
1000	Complete		0.987	0.987	0.992	0.979
	30%		0.956	0.959	0.970	0.972
	50%		0.898	0.899	0.931	0.951
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	1.000
	50%		0.998	0.998	0.999	1.000
400	Complete	nonuniform	0.938	0.940	0.922	0.794
	30%		0.897	0.902	0.902	0.590
	50%		0.769	0.777	0.813	0.424
1000	Complete		0.997	0.997	0.994	0.953
	30%		0.995	0.994	0.995	0.758
	50%		0.983	0.984	0.990	0.517
2000	Complete		1.000	1.000	1.000	1.000
	30%		1.000	1.000	1.000	0.914
	50%		1.000	1.000	1.000	0.541

Table C8.

Item level Type I error rates of methods in thresholds non-invariant conditions

Sample size	Missing rate	Type of DIF	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
			α = 0.01/cutoff MI: 6.635		α = 0.05/cutoff MI: 3.841	
400	Complete	small	0.007	0.011	0.047	0.053
	30%		0.007	0.012	0.045	0.054
	50%		0.008	0.017	0.046	0.070
1000	Complete		0.012	0.014	0.060	0.060
	30%		0.012	0.020	0.061	0.075
	50%		0.011	0.037	0.060	0.115
2000	Complete		0.013	0.013	0.052	0.059
	30%		0.014	0.028	0.050	0.091
	50%		0.012	0.083	0.050	0.171
400	Complete	large	0.012	0.013	0.057	0.058
	30%		0.012	0.014	0.057	0.059
	50%		0.011	0.017	0.056	0.073
1000	Complete		0.008	0.011	0.047	0.054
	30%		0.009	0.015	0.046	0.066
	50%		0.009	0.034	0.046	0.106
2000	Complete		0.009	0.013	0.052	0.061
	30%		0.009	0.029	0.052	0.094
	50%		0.009	0.088	0.052	0.174
400	Complete	mixed	0.008	0.014	0.047	0.058
	30%		0.009	0.014	0.046	0.056
	50%		0.008	0.016	0.046	0.071
1000	Complete		0.006	0.013	0.047	0.055
	30%		0.007	0.019	0.047	0.070
	50%		0.007	0.036	0.047	0.107
2000	Complete		0.009	0.014	0.041	0.059
	30%		0.009	0.028	0.041	0.084
	50%		0.008	0.085	0.042	0.168
400	Complete	nonuniform	0.009	0.011	0.047	0.050
	30%		0.009	0.011	0.047	0.050
	50%		0.009	0.012	0.047	0.062
1000	Complete		0.009	0.012	0.050	0.053
	30%		0.010	0.029	0.050	0.102
	50%		0.012	0.012	0.059	0.055
2000	Complete		0.013	0.027	0.058	0.091
	30%		0.012	0.083	0.057	0.167
	50%		0.010	0.029	0.050	0.102

Table C9.

Item level power of methods in thresholds non-invariant conditions

Sample size	Missing rate	Type of DIF	WLSMVMI	WLSMVPD	WLSMVMI	WLSMVPD
			α = 0.01/cutoff MI: 6.635		α = 0.05/cutoff MI: 3.841	
400	Complete	small	0.088	0.146	0.212	0.354
	30%		0.079	0.167	0.202	0.435
	50%		0.059	0.188	0.161	0.492
1000	Complete		0.223	0.458	0.431	0.702
	30%		0.187	0.576	0.401	0.825
	50%		0.142	0.708	0.316	0.912
2000	Complete		0.425	0.863	0.662	0.951
	30%		0.368	0.959	0.613	0.990
	50%		0.299	0.992	0.525	0.999
400	Complete	large	0.384	0.781	0.637	0.913
	30%		0.323	0.730	0.587	0.904
	50%		0.243	0.655	0.483	0.909
1000	Complete		0.784	0.991	0.898	0.996
	30%		0.694	0.992	0.875	0.999
	50%		0.561	0.991	0.777	1.000
2000	Complete		0.981	1.000	0.998	1.000
	30%		0.962	1.000	0.994	1.000
	50%		0.896	1.000	0.971	1.000
400	Complete	mixed	0.421	0.683	0.619	0.844
	30%		0.360	0.665	0.573	0.850
	50%		0.276	0.600	0.471	0.862
1000	Complete		0.711	0.943	0.856	0.974
	30%		0.653	0.955	0.830	0.988
	50%		0.551	0.966	0.752	0.995
2000	Complete		0.932	0.999	0.980	1.000
	30%		0.891	1.000	0.957	1.000
	50%		0.831	1.000	0.933	1.000
400	Complete	nonuniform	0.316	0.711	0.519	0.876
	30%		0.274	0.491	0.491	0.668
	50%		0.221	0.375	0.408	0.498
1000	Complete		0.613	0.967	0.794	0.987
	30%		0.463	0.521	0.695	0.570
	50%		0.906	1.000	0.969	1.000
2000	Complete		0.878	0.932	0.964	0.969
	30%		0.808	0.550	0.930	0.593
	50%		0.463	0.521	0.695	0.570